Statistics Surveys
Vol. 16 (2022) 210–237
ISSN: 1935-7516
https://doi.org/10.1214/22-SS138

Central subspaces review: methods and applications

Sabrina A. Rodrigues

Department of Epidemiology and Biostatistics, School of Public Health
Imperial College London, London UK
e-mail: s.rodrigues@imperial.ac.uk

Richard Huggins

School of Mathematics and Statistics, The University of Melbourne, Australia
e-mail: rhuggins@unimelb.edu.au

and

Benoit Liquet

Laboratoire de Mathématiques et de leurs Applications de Pau, Université de Pau et des Pays de l'Adour, Pau, France

School of Mathematical and Physical Sciences, Macquarie University, Sydney, New South Wales, Australia
e-mail: benoit.liquet-weiland@mq.edu.au

Abstract: Central subspaces have long been a key concept for sufficient dimension reduction. Initially constructed for solving problems in the $p < n$ setting, central subspace methods have seen many successes and developments. However, over the last few years and with the advancement of technology, many statistical problems are now situated in the high dimensional setting where $p > n$. In this article we review the theory of central subspaces and give an updated overview of central subspace methods for the $p \leq n$, $p > n$ and big data settings. We also develop a new classification system for these techniques and list some R and MATLAB packages that can be used for estimating the central subspace. Finally, we develop a central subspace framework for bioinformatics applications and show, using two distinct data sets, how this framework can be applied in practice.

MSC2020 subject classifications: 62P10, 92B15.
Keywords and phrases: Central subspaces, sufficient dimension reduction, dimension reduction subspaces, sliced inverse regression, bioinformatics, omics.

Received April 2021.

Contents

1. Introduction

The term *central subspace* was first used by Cook (1994a) but it is a concept that has been built upon previous ideas of Li (1991) and Cook (1994b). In his seminal paper, Li (1991) established a new line of research by developing the Sliced Inverse Regression (SIR) technique, which aims to determine a specific dimension reduction space, the *effective dimension reduction space* (EDR). In subsequent years, Cook (1994a, 1998) reinterpreted and extended Li's idea by interpreting the EDR in terms of a *dimension reduction subspace* (DRS)[1] and introducing the concept of *sufficient dimension reduction* (SDR) (Cook, 1994a, 1998, 2007; Cook, Li and Chiaromonte, 2007). He further combined ideas of conditional independence of Dawid (1979) and addressed issues of existence and uniqueness to develop the current theory of DRS. Through this development, the concepts of *minimum dimension reduction subspace* and *central dimension reduction subspace*, also known as *central subspace*, emerged.

This paper is organised as follows: in Section 2 we review the theory of central subspaces and dimension reduction subspaces; in Section 3 we provide a literature review of some methods that aim to estimate the central subspace, introduce a new classification scheme for central subspace methods and provide a list of R and MATLAB packages that can be used for estimation of the central subspace; in Section 4 we provide a framework that shows how central subspaces can be applied to bioinformatics and demonstrate these ideas using two real datasets; finally, in Section 5 we provide a summary and discuss some practical issues. Note that we do not discuss estimating the dimension of the central subspace here as a detailed review is provided in Ma and Zhu (2013).

[1] In the literature these two terms - EDR and DRS - are often confused. It is important to note that they are posed in different settings but their definitions are equivalent, as shown in Zeng and Zhu (2010).

2. Definitions and properties of the central subspace

Central subspaces are a particular case of dimension reduction subspaces. In this section we first review some definitions and properties and then provide a formal definition. Our exposition follows that of Cook (2018) and Chapter 6 of Cook (1998), where proofs and further details can be found.

Definition 2.1 (Dimension Reduction Subspace)**.** Let $B = (b_1, \ldots, b_d)$ denote a $p \times d$ matrix with $d \leq p$ linearly independent $p \times 1$ columns $b_1, b_2, \ldots, b_d$, let Y denote the response variable and let $X = (X_1, \ldots, X_p)^T$ be the $p \times 1$ predictor vector such that

$$F_{Y|X}(y|x) \quad = \quad F_{Y|b_1^T X, b_2^T X, \ldots, b_d^T X}(y|b_1^T x, b_2^T x, \ldots, b_d^T x) \tag{1}$$

Then the span of the columns of B, denoted $S(B)$, is called a *dimension reduction subspace (DRS)* for $Y|X$, or equivalently, for the regression of Y on X.

Expression (1) is equivalent to saying that X and Y are independent given $B^T X$ (Basu and Pereira, 1983), that is,

$$Y \perp\!\!\!\perp X|B^T X \tag{2}$$

In other words, Y only depends on X through $B^T X$ and there would be no loss of information if X were replaced by $B^T X$. Thus, if (2) holds, Y and $B^T X$ are sufficient (Cook, 1994a, 1998, 2007; Cook, Li and Chiaromonte, 2007) to determine the relationship between Y and X.

Notice that in Definition 2.1 if $B = I_p$, where I_p is the $p \times p$ identity matrix, the DRS may not result in reduced dimensionality since, in this case, the dimension reduction subspace coincides with the original subspace. Furthermore, if we replace $B^T X$ with $A^T X$ where $A \neq B$ but such that $S(A) = S(B)$, the equality in expressions (1) and (2) still hold. This leads us to two conclusions:

(I) DRSs need not be unique
(II) If $S(A)$ and $S(B)$ are two DRSs then we can have $\dim(S(A)) \leq \dim(S(B))$.

Conditions (I) and (II) imply the existence of different subsets of predictors spanning a DRS but which might not agree in number, that is, one set can be "smaller" than the other. This leads us to the next definition.

Definition 2.2 (Minimum Dimension Reduction Subspace)**.** A subspace S is said to be a *minimum DRS* for $Y|X$ if the following two conditions hold:

i) S is a DRS for $Y|X$; and
ii) $\dim(S) \leq \dim(S')$ for all DRS S'

Therefore, the minimum dimension reduction subspace contains the minimum number of predictors, or linear combinations of predictors, required for equations (1) and (2) to hold. As noted by Cook (1998), Definition 2.2 does not guarantee the uniqueness of the minimum DRS but guarantees the minimum dimension of a DRS.

We are interested in obtaining a unique minimum DRS, therefore we have to further constrain Definition 2.2 to define such a DRS. This leads us to the key definition:

Definition 2.3 (Central Dimension Reduction Subspace)**.** A subspace S is a *central dimension reduction subspace* for the regression of Y on X, if S is a DRS and $S \subset S'$, for all DRS S'.

We refer to the central dimension reduction subspace for the regression of Y on X as the *central subspace* and we denote it by $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{Y|X}(B)$ when the basis is not explicit.

From the definition of central subspace we have the following proposition:

Proposition 2.4. *The central subspace, $\mathcal{S}_{Y|X}$, exists if and only if the following two conditions hold:*

(i) the intersection, $\cap S'$, of all DRS S' is itself a DRS; and,
(ii) $\mathcal{S}_{Y|X} = \cap S'$

Proposition 2.4 tells us that the central subspace exists only when the intersection of all DRS is itself a DRS and in that case it is the intersection of all DRS. However, note that the intersection of DRS always exists, but it might not be a DRS. See Chapter 6 of Cook and Weisberg (1999) for examples.

Although we can not guarantee the existence of the central subspace we can assure that when it exists it is unique and is the subspace with mininum dimension for the regression of Y on X. This is guaranteed by the following proposition:

Proposition 2.5. *If $\mathcal{S}_{Y|X}$ is the central dimension reduction subspace for the regression of Y on X, then $\mathcal{S}_{Y|X}$ is the unique minimum dimension reduction subspace.*

Note that a unique minimum DRS needs not be a central subspace. This occurs when the central subspace does not exist, that is, when the intersection of DRS is not a DRS. To see this clearly, consider the following example:

Example 2.1. *Let $p = 3$ and let $X = (X_1, X_2, X_3)$ be uniformly distributed on the unit sphere where $||X|| = 1$ and set $Y|X = X_1^2 + \varepsilon$, where ε is an independent error. The subspace $S(\{(1,0,0)^T\})$ is the unique minimum DRS and the subspace $S'(\{(0,1,0)^T, (0,0,1)^T)\})$ is a DRS. However, S is not contained in S' and their intersection is the origin which is not a DRS. Hence the central subspace in this case does not exist.*

Despite the difficulty of determining and guaranteeing the existence of the central subspace, there are conditions under which the central subspace exists. For further details see Chapter 6 of Cook (1998). Throughout this paper we assume their existence.

3. Techniques to estimate the central subspace

The literature on techniques to estimate the central subspace is extensive and growing. The publication of Li's seminal paper (Li, 1991) on Sliced Inverse Regression (SIR) brought not only a new concept to the field of dimension reduction but it also initiated a new line of research, leading to an increasing variety of techniques. In an attempt to overcome the limitations of SIR in determining the central subspace, other researchers have developed improved versions of SIR or developed techniques that consider the problem from another viewpoint. Cook (1998) considers numerical and graphical techniques to estimate the central subspace. Here, we consider only numerical techniques and focus in more detail on those that have extended the methods of estimating the central subspace to the high dimensional setting. For a thorough review on some of the lower-dimensional techniques see for example, Li, Zha and Chiaromonte (2005), Xia (2007) Fukumizu, Bach and Jordan (2009), Ma and Zhu (2012), Cook, Forzani and Rothman (2012). For a good compendium of the graphical techniques see Cook (1998); for a detailed review with many examples applied to real datasets with R see Li (2018) and for more recent and advanced topics such as multivariate response SIR and variable selection in SIR see Girard, Lorenzo and Saracco (2022).

3.1. Sliced inverse regression

Slice Inverse Regression (SIR) (Li, 1991; Cook, 1998), as opposed to traditional regression techniques such as Ordinary Least Squares, performs the regression of X on Y, benefiting from Y being, usually, of lower dimension than X. Generally speaking, SIR slices the response variable into h slices; calculates $\mathbb{E}(X|Y)$ on each slice; and then performs principal component analysis using the calculated means of each slice to determine the largest eigenvectors that are associated with the subspace of interest.

In the literature SIR is criticized for:

(I) Depending on the linearity condition. SIR assumes $\mathbb{E}(X|B^TX)$ to be a linear function of B^TX, that is, for any $b \in \mathbb{R}^p$

$$\begin{aligned}\mathbb{E}(b^TX|B^TX) &= \mathbb{E}(b^TX|b_1^TX, b_2^TX, \ldots, b_d^TX)\\ &= c_0 + c_1b_1^TX + \ldots + c_db_d^TX\end{aligned}$$

for some scalars $c_0, c_1, \ldots, c_d \in \mathbb{R}$. This condition can be shown to be a characterization of elliptical symmetry of the distribution of X (Cook and Weisberg, 1991). However, this assumption does not always hold in practice but as mentioned in Chen and Li (1998) and Prendergast (2007), Hall and Li (1993) have shown this condition to hold approximately in the presence of high-dimensional data.

(II) Failing to estimate the central subspace under some conditions; or, performing a non-exhaustive search of the central subspace directions, that is, SIR might not always determine vectors that are in the central subspace. See Section 3.2 for further details on this point.

(III) Not making use of the characterization of conditional independence given in Definition 2.1. Although it does estimate dimension reduction directions, as mentioned in Xia (2007), the fact that it does not use Definition 2.1 leads to its failing to efficiently estimate the central subspace.

(IV) Requiring the inverse of the sample covariance matrix. This prevents it from being applied in high dimensional settings where $p \gg n$. Cook, Forzani and Rothman (2012) further add that most post-SIR SDR regression techniques also suffer from this problem.

(V) Estimating a linear subspace and thus being unable to find nonlinear features (Wu, 2008). This is an issue as there is no guarantee that the central subspace will be linear.[2]

SIR is a very ingenious idea and constitutes one of the great advancements for the field of dimension reduction. In the next section we show how many central subspace techniques, have emerged from trying to overcome some of these limitations. Note that there are other lines of research considering SIR, such as, its robustness analysis to influential outliers. We do not discuss these here as it is not in the scope of this paper. For further details on this topic see Prendergast (2005), Prendergast (2007) and Prendergast and Smith (2010).

3.2. A glimpse at the last 30 years

Early approaches addressed improving SIR by including additional assumptions and by using different methods to calculate the directions of the subspace of interest. Of particular concern was the issue of SIR not being able to recover the central subspace under certain conditions, namely, if the regression function is symmetric about 0. As stated in Li (2018), if the regression surface is symmetric about 0 along some directions, then those directions cannot be recovered by central subspace methods that rely soley on first-order conditional moments (see Section 3.2 of Li (2018) for an example). Cook and Weisberg (1991) and Cook (1998) suggest that without a further restriction – $\text{Cov}(X|B^TX)$ be a constant matrix – SIR might determine the directions that fall in a dimension reduction subspace but these might not span the subspace and they need not belong to the central subspace. This exploration of SIR's shortcomings led then to methods such as SAVE (Cook and Weisberg, 1991), LAD (Cook and Forzani, 2009) and others. The SAVE method uses averages of the variance to estimate the dimension reduction directions as opposed to averages of the means; pHd (Li, 1992), estimates the directions by using principal Hessian directions; CANCOR (Fung et al., 2002), similar to SIR and SAVE, uses a B-spline method to estimate $\text{Cov}(\mathbb{E}(X|Y))$; KIR (Zhu and Fang, 1996), uses a kernel based estimate for $\text{Cov}(\mathbb{E}(X|Y))$; and, other methods such as CORE (Cook and Forzani, 2008a), LAD (Cook and Forzani, 2009)) and PFC (Cook and Forzani, 2008b))

[2] The term subspace, by definition, is associated to a linear space. However, here we also use it to refer to a subspace with nonlinear features.

adopt a likelihood-based approach. Most of these techniques still rely on one or both of the conditions on moments to estimate the directions efficiently, which leads to assuming elliptical distributions of the covariate vector X. pHd estimates directions of the central mean subspace which is contained in the central subspace and is effective for the regression of $\mathbb{E}(Y|X)$ (Fukumizu, Bach and Jordan, 2009). LAD, CORE and PFC assume the conditional distribution of X given the response Y to be multivariate normal.

Another approach, developed by Cook and Nachtsheim (1994), circumvents the restriction of SIR for elliptically distributed covariates – condition (I) – by transforming or reweighting the predictors. More recently, Li and Dong (2009) and Dong and Li (2010) proposed an inverse regression framework that removes the linearity condition and generalizes several moment-based inverse regression methods. However, they still require the constant variance condition.

Other methods evolved from noting that SIR does not use the characterization of conditional independence stated in the initial problem (condition III) and does not detect the directions exhaustively (condition II). This led to methods such as KDR (Fukumizu, Bach and Jordan, 2004, 2009), a Fourier transform method suggested by Zhu and Zeng (2006) and dMAVE (Xia, 2007). KDR uses a characterization of conditional independence through covariance operators. Using this characterization, they design an objective function which is minimized when the conditional independence assertion in Definition 2.1 is realized. KDR estimates a linear subspace and requires the optimization of a nonconvex function (Kim and Pavlovic, 2011), which in general is more computationally demanding than convex problems. The Fourier transform method suggested by Zhu and Zeng (2006), is similar to KDR in its characterization of conditional independence through the use of Hilbert spaces and an objective function but it requires estimating the derivative of the density of X. As mentioned by Fukumizu, Bach and Jordan (2009), in practice this places a normality restriction on X, which again might not hold. The dMAVE method of Xia (2007) estimates the conditional density function of $Y|B^TX$ and then proposes two distinct algorithms to detect the directions of the central subspace. As noted by Ma and Zhu (2012), it requires estimating the distribution function non-parametrically and requires continuous covariates.

Another line of research focused on SIR, and alternative methods, being limited to the estimation of a linear dimension reduction subspace (condition V). Hence nonlinear methods such as mKDR (Nilsson, Sha and Jordan, 2007), KSIR (Wu, 2008) and COIR (Kim and Pavlovic, 2011) emerged. KSIR and mKDR are nonlinear extensions of SIR and KDR, respectively, and thus also suffer from their drawbacks. COIR combines ideas from KDR and SIR. It estimates the variance of the inverse regression method but avoids slicing the response variable by using covariance operators in reproducing kernel Hilbert spaces. According to Kim and Pavlovic (2011), COIR outperforms previous methods on many real world problems. However, from a close analysis, COIR, mKDR and KDR require selecting kernel parameters to obtain good estimates of the directions of the central subspace. This involves methods such as cross-validation that carry an extra computational burden. For a general theory on non-linear sufficient dimension reduction see Lee, Li and Chiaromonte (2013).

Most of the methods mentioned above have been developed in the $p \leq n$ setting. However, many datasets and current real world problems fall into the $p > n$ setting. As mentioned in Cook, Forzani and Rothman (2012), nearly all SDR methods available require the inverse of a $p \times p$ sample covariance matrix, which leads to problems when $p > n$. One method of overcoming this problem is to avoid the computation of the inverse of the sample covariance matrix altogether, as seen with the seeded dimension-reduction method (Cook, Li and Chiaromonte, 2007). However, as stated by the authors, the methodology can only be used when n is large relative to another parameter u, which lies in between d and p. Moreover, the authors emphasize that in its current version, their methodology cannot be used in the "large p small n" regressions. Another approach has been to replace the p predictors with $p^* \ll n$ principal components (linear combinations of the predictors) and then apply an SDR method to the regression of the response on the selected p^* predictors. Examples of these include Chiaromonte and Martinelli (2002) and Li and Li (2004). However, as noted in Cook, Forzani and Rothman (2012), recent results on the eigenvectors of sample covariance matrices in high-dimensional settings raise questions on this approach.

Other alternate directions of extending SIR to the high dimensional setting when $p \gg n$ are those of performing regularization, variable selection and using kernel methods. Regularized Sliced Inverse Regression (RSIR) (Zhong et al., 2005) and the Sliced Inverse Regression method based on the QZ algorithm (SIR-QZ) (Coudret, Liquet and Saracco, 2014) are examples of regularization based central subspace methods. RSIR determines a bootstrap estimator of the mean squared error of regularized estimates of the EDR directions. The SIR-QZ method estimates the indices rather than the EDR directions by making use of several estimations from the various number of slices and by trying to find a minimial regularization of the covariance matrix of the predictors. For further details on these regularized approaches to SIR see Coudret, Liquet and Saracco (2014) and Girard, Lorenzo and Saracco (2022) for a review, and Bernard-Michel, Gardes and Girard (2008) which identify theoretical problems with the ridge SIR estimator defined by Li and Yin (2008) and propose an alternative ridge SIR estimator. Note that some of these regularized approaches although successful create orthogonal basis vectors of the central subspace that are linear combinations of all the predictor variables and consequently can be time consuming to run in the case where p is large.

Methods employing variable selection, on the other hand, assume that only a few variables are relevant for explaining and predicting the response variable. This sparsity assumption has led to what is known as *Spare Sufficient Dimension Reduction* and dates back to Cook (2004), Li and Nachtsheim (2006) and Li (2007), as mentioned by Lin, Zhao and Liu (2018) and Li (2018). Some of the methods proposed combine SIR based methods with variable selection methods involving regularization algorithms such as those available in the Lasso (Tibshirani, 1996), the Elastic-net (Zou and Hastie, 2005) and the Dantzig selector (Candes and Tao, 2007). These combined methods provide a sparse linear combination for the central subspace basis vectors, that is, the basis vectors are no longer a linear combination of all of the predictors (as in RSIR or SIR-QZ, for example) but are a linear combination of the variables that have been selected. Examples of these include: Sparse Ridge Sliced Inverse Regression (SR-SIR) (Li and Yin, 2008), Diagonal Threshold Screening SIR (DT-SIR) (Lin, Zhao and Liu, 2018) and more recently the Lasso-SIR (Lin, Zhao and Liu, 2019). Li and Yin (2008) applied their method to a dataset where $n = 240$ and $p = 7399$. However, they first pre-selected 329 of those predictors to speed up the computation, suggesting that if p is much larger than n, computation speed is compromised. Lin, Zhao and Liu (2019) has successfully applied Lasso-SIR to a dataset with $p = 47293$ and $n = 90$ without performing any pre-selection. The computational efficiency of the Lasso-SIR is due in part to their method being reliant on the computational efficient algorithms developed in the glmnet package (Simon et al., 2011). For a review of sparse sufficient dimension reduction methods see Li (2018, section 15.4), Lin, Zhao and Liu (2019), Girard, Lorenzo and Saracco (2022), Nghiem et al. (2023) for more recent developments and Qian, Ding and Cook (2019). The latter, in addition to reviewing the sparse based methods also propose a unified solution involving a minimum discrepancy approach with regularization which allows methods such as SIR (Li, 1991), PFC (Cook and Forzani, 2008b) and SAVE (Cook and Weisberg, 1991) to be applied in the setting where $p \gg n$. For theoretical concerns of the sparse based SIR estimators, such as, consistency and optimal rates of convergence, see Lin, Zhao and Liu (2018) and Tan, Shi and Yu (2020), respectively.

Regarding the use of kernel methods, Fukumizu and Leng (2014) use the KDR method of Fukumizu, Bach and Jordan (2004, 2009) to propose a new gradient-based KDR estimator, gKDR. gKDR estimates the gradient of the regression function, allowing them to avoid the optimization problem encountered in Fukumizu, Bach and Jordan (2009) that prevents it from being applied directly to a high-dimensional setting. The authors show that gKDR can be applied to low and high dimensional settings and achieves competitive or improved results when compared to other techniques. Nevertheless in our view it has two strong limitations. First, and as mentioned in Fukumizu and Leng (2014), when applied to classification problems with L classes, gKDR can only find central subspaces with dimension d of at most $L-1$. Note that this problem also affects SIR based methods. To overcome this difficulty Fukumizu and Leng (2014) developed gKDR-v which involves partitioning the samples into subsets and deriving the gKDR-v estimator from these subsets. The second difficulty affects all variants of gKDR (gKDR, gKDR-i and gKDR-v). gKDR methods use cross-validation, paired with k-nearest neighbours to estimate the dimension of the central subspace. This can be problematic as it imposes an extra computational burden.

A somewhat different approach for applying central subspace methods to the $p > n$ setting is that of Yin and Hilafu (2015). The authors propose a general sequential dimension reduction framework that allows methods developed in the $p < n$ setting to be applied to the $p > n$ setting. The framework partitions the predictor vector into smaller subvectors such that $p < n$ and then performs reduction on those subvectors sequentially, thus avoiding the inversion of the $p \times p$ sample covariance matrix. However, as noted by the authors, differences between solutions from different partitions of the predictor vector may arise and may be difficult to quantify. Furthermore, and as mentioned by Hilafu and Yin (2017), this approach is computationally intensive as the tuning parameters for both the ℓ_1 and ℓ_2 norm penalties need to be chosen at each step of the sequential reduction framework and their algorithms cannot handle the case where predictors exhibit high correlations. To deal with the latter, Hilafu and Yin (2017) developed the sequential partial inverse regression (SeqPIR) method, which takes advantage of ideas from partial least squares, dimension reduction in regression without matrix inversion (Cook, Li and Chiaromonte, 2007) and the work of Yin and Hilafu (2015).

With respect to central subsapce methods being applied to massive datasets[3] such as those often encountered in the big data setting, Liquet and Saracco (2016) have developed BIG-SIR. To our knowledge, the authors are the first to apply SIR to a massive dataset. The method is based on the "divide and conquer" or "divide and recombine" principle (Lin and Xi, 2011; Guha et al., 2012; Chen and Xie, 2014). This principle consists of splitting the data over n into K blocks, applying the method on each block and then aggregating the K results to produce an analysis of the complete data. Although constituting an excellent advancement for SIR based methods to massive datasets, BIG-SIR is only applicable to studies where $p < n$, and in the case of an L-classification problem, can only estimate $L - 1$ dimensions.

3.3. A new classification system and software packages for central subspace methods

As mentioned previously, Cook (1998) and Cook and Weisberg (1999) initially suggested that we could distinguish between two main types of techniques to estimate the central subspace: graphical and numerical. Additionally, he separated numerical techniques into *standard fitting methods* and *inverse regression methods*. This excludes novel techniques such as KDR (Fukumizu, Bach and Jordan, 2004, 2009) and mKDR (Nilsson, Sha and Jordan, 2007), since these are not standard nor inverse regression techniques. There are many other classifications in the literature such as *semi-parametric* versus *non-parametric* (Fukumizu, Bach and Jordan, 2004) and *direct regression* versus *inverse regression* estimation methods (Xia, 2007). However, we find that the first leads to ambiguity since, for example, the Fourier method suggested by Zhu and Zeng (2006) is considered to be non-parametric but requires assuming parametric models for X (Fukumizu, Bach and Jordan, 2009). The classification given by Xia (2007) leaves his own method unclassified as it uses ideas both from inverse regression and direct regression.

[3] Emerson and Kane (2012) define a dataset to be large if it exceeds 20% of the RAM on a machine and massive if it exceeds 50%.

More recently, Cook, Forzani and Rothman (2012) suggest that *this body of work reflects three different but related frontiers: extensions that require progressively fewer assumptions, development of likelihood based methods and adaptations for specific areas of application.* To incorporate high-dimensional data applications, we propose to classify central subspace methods according to the following criteria: type of **regression**, inverse or non-inverse regression; type of **subspace** determined, linear or non-linear; **dimensions** of the data it can be applied to, $p \leq n$ or $p > n$; and,**size**, regular or big data.

The criteria are self-explanatory but briefly, we consider inverse regression (IR) methods, methods that aim to estimate the central subspace using inverse regression. These would include, for example, all SIR based methods. Non-inverse regression (NIR) methods are those that either use standard forward regression, a combination of inverse-regression and forward regression or other methods. We consider a central subspace method non-linear (NL) if it allows the estimation of a non-linear central subspace and is linear (L) if it does not. A central subspace method is categorised as $p > n$ if it can be applied in the setting where $p > n$; it is $p \leq n$ otherwise. Note that a method that can be applied in the $p > n$ setting can also be applied in the $n < p$ setting. To include large/massive central subspace methods in our classification we define a central subspace method as *big-data (BD) central subspace method* if it can be applied to large or massive datasets in the Emerson and Kane (2012) sense; and, regular data (RD) otherwise. Table 1 presents the classification of some of the aforementioned methods regarding the type of regression, subspace determined, dimension setting and size of data they can be applied to. Table 1 focuses mostly on SIR based methods but as seen in the table and mentioned here within, non-SIR based methods exist. Table 1 demonstrates how we can use our classification system to characterize central subspace methods according to the four criteria presented.

In Table 2 we list available R and MATLAB packages that implement methods which aim to estimate the central subspace. Note that some of the methods presented in the table have thus far not been discussed. For further information on these see the respective references contained within the table.

4. A central subspace framework for bioinformatics

In this section we provide a framework for the use of central subspaces in bioinformatics or omics applications. For demonstration purposes, we consider gene expression studies and explore these in a semi-informal way through some examples and apply the framework to a real dataset in Section 4.4. We focus on this application in bioinformatics but note that it can be applied to other applications such as those involving SNP data, protein expression data, metabolomic data, RNA-seq data or ultimately a combination of several different types of data and a multivariate response. We use the expression "determine the central subspace of a disease" to mean "determine the vectors that form a basis for the central subspace of $Y|X$, where Y is the response variable for a disease and X is

TABLE 1
Classification of central subspace estimation methods.

Method	Reference	Regression	Subspace	Dimension	Size
SIR	Li (1991)	IR	L	$p \leq n$	RD
SAVE	Cook and Weisberg (1991)	IR	L	$p \leq n$	RD
KIR	Zhu and Fang (1996)	IR	L	$p \leq n$	RD
CANCOR	Fung et al. (2002)	IR	L	$p \leq n$	RD
MAVE	Xia et al. (2002)	NIR	L	$p \leq n$	RD
KDR	Fukumizu, Bach and Jordan (2004, 2009)	NIR	L	$p \leq n$	RD
RSIR	Zhong et al. (2005)	IR	L	$p > n$	RD
KSIR	Wu (2008)	IR	NL	$p > n$	RD
SR-SIR	Li and Yin (2008)	IR	L	$p > n$	RD
	Bernard-Michel, Gardes and Girard (2008)				
COIR	Kim and Pavlovic (2011)	IR	NL	$p > n$	RD
SIR-QZ	Coudret, Liquet and Saracco (2014)	IR	L	$p > n$	RD
gKDR	Fukumizu and Leng (2014)	NIR	L	$p > n$	RD
SeqPIR	Hilafu and Yin (2017)	IR	L	$p > n$	RD
DT-SIR	Lin, Zhao and Liu (2018)	IR	L	$p > n$	RD
Lasso-SIR	Lin, Zhao and Liu (2019)	IR	L	$p > n$	RD
BIG-SIR	Liquet and Saracco (2016)	IR	L	$p < n$	BD

the vector of predictors". Thus, our goal for this section is to show how central subspaces can be used to determine the *central subspace of a disease*, that is, the smallest set of predictors that is involved in determining the response variable, which in theory can be also be multivariate.

4.1. Background

We wish to determine the smallest dimension reduction space for which we can make inferences about the conditional distribution of $Y|X$. In theory, it is the central subspace that we are interested in but in real world applications, for now, we are interested in determining the vectors that form a basis for the central subspace. We often start with $p \gg n$ predictors and by determining the vectors in the basis for the central subspace, we reduce the dimensionality of the problem. By determining its dimension we will be able to make inferences on some of the distinct combinations that may lead to the response variable Y. Once these basis vectors are determined we can use them to establish a relationship between the distribution of Y and some function of the basis vectors.

Let Y and $X = (X_1, \dots, X_p)^T$ denote respectively, the random variables that correspond to the response and the predictor vector. For example, X_j, $j = 1, \dots, p$ may be the expression measure for gene j. Suppose we can find a $p \times d$ matrix $B = (b_1, \dots, b_d)$ that satisfies Definition 2.1 and subsequently (2), that is,

$$Y \perp\!\!\!\perp X | B^T X$$

TABLE 2
List of some software packages for methods that aim to determine the central subspace.

Software	Source	Package	Method
R	CRAN	dr Weisberg (2002)	SIR Li (1991)
			SAVE Cook and Weisberg (1991)
			pHdy Li (1992); Cook (1998)
			pHdres Cook (1998)
			pHdq Li (1992)
			IRE Cook and Ni (2005)
	CRAN	ldr[1] Adragni and Raim (2014)	CORE Cook and Forzani (2008a)
			LAD Cook and Forzani (2009)
			PFC Cook (2007); Cook and Forzani (2008b)
	CRAN	edrGraphicalTools Liquet and Saracco (2011)	SIR Li (1991)
			SAVE Cook and Weisberg (1991)
			RSIR Zhong et al. (2005)
			SR-SIR Li and Yin (2008); Bernard-Michel, Gardes and Girard (2008)
			SIR-QZ Coudret, Liquet and Saracco (2014)
	CRAN	MAVE Hang and Xia (2017)	KSIR Wang and Xia (2008)
			MAVE Xia et al. (2002)
			OPG Xia (2007)
	CRAN	LassoSIR Zhao, Lin and Liu (2017)	Lasso-SIR Lin, Zhao and Liu (2019)
MATLAB	Journal website	LDR[2] Cook, Forzani and Tomassi (2011)	CORE Cook and Forzani (2008a)
			LAD Cook and Forzani (2009)
			PFC Cook (2007); Cook and Forzani (2008b)
	Author's webpage	KSIR	KSIR Wu (2008)
	Author's webpage	KDR	KDR Fukumizu, Bach and Jordan (2004, 2009)
	Author's webpage	gKDR	gKDR Fukumizu and Leng (2014)

[1] Analagous to the LDR MATLAB package
[2] Analagous to the ldr R package

Now,

$$
\begin{aligned}
B^T X &= \begin{bmatrix} b_{11} & b_{21} & \dots & b_{p1} \\ b_{12} & b_{22} & \dots & b_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1d} & b_{2d} & \dots & b_{pd} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \\
&= \begin{bmatrix} b_{11}X_1 + b_{21}X_2 + \ldots + b_{p1}X_p & \dots & b_{1d}X_1 + b_{2d}X_2 + \ldots + b_{pd}X_p \end{bmatrix}^T \\
&= \left(\sum_{i=1}^{p} b_{i1}X_i, \sum_{i=1}^{p} b_{i2}X_i, \ldots, \sum_{i=1}^{p} b_{id}X_i \right)^T
\end{aligned}
$$

$$= \left(b_1^T X, b_2^T X, \ldots, b_d^T X\right)^T$$

By observing the equations above and knowing that B satisfies (2) we can see that, in essence, B expresses, through its columns, the different combinations that are associated with Y. In other words, by determining B we are determining the vectors $b_1, \ldots, b_d$ that, when applied to X, give us the linear combinations of the explanatory vectors that are required to determine the conditional distribution of the response Y, through some function f. In our gene expression example, this means that the matrix B returns the distinct combinations of genes that are involved in the expression of the response Y. To illustrate consider two main types of diseases mentioned in Lvovs, Favorova and Favorov (2012), monogenic and polygenic.

4.2. Central subspace of a mongenic disease

Monogenic diseases, also known as Mendelian diseases, are caused by a single mutation in one gene (Lvovs, Favorova and Favorov, 2012). Well known examples include sickle cell anaemia, cystic fibrosis and Huntington's disease.[4]

Let Y be the response variable for a certain monogenic disease D caused by a gene, which we call gene 2, represented by X_2. Here Y is either 1 or 0, according to whether the individual has the disease or not. Then

$$Y \sim \text{bern}(f(X_2)),$$

where bern denotes the Bernoulli distribution and f is a real valued bounded function in $[0, 1]$. In this case, the central subspace has dimension 1 and is spanned by the vector b_1, which is a p vector with $p - 1$ zeros and which at position 2 has b_{21}, that is, $b_1 = (0, b_{21}, 0, \ldots, 0)^T$. Using matrix notation we have

$$\begin{aligned} B^T X &= \begin{bmatrix} 0 & b_{21} & \ldots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \\ &= 0 \times X_1 + 1 \times b_{21} X_2 + \ldots + 0 \times X_p \\ &= b_{21} X_2 \end{aligned}$$

Thus, in the case of a monogenic disease caused by gene i, the central subspace has dimension one and a basis for it will be $\{b_1\}$ with $b_1 = (0, \ldots, b_{i1}, \ldots, 0)^T$ and $b_{i1} \neq 0$, $b_{i1} \in \mathbb{R}$. Note that Y could also be continuous, for example, blood pressure, expression of a particular gene, etc.

[4] http://www.who.int/genomics/public/geneticdiseases/en/index2.html

4.3. Central subspace of a polygenic disease

Polygenic diseases or complex diseases are caused by mutations in more than one gene (Lvovs, Favorova and Favorov, 2012). We can have several cases:

> **Case 1:** The disease arises from a particular combination of genes. We call this a *simple polygenic disease*.
> **Case 2:** The disease may arise from different combinations of genes. We call this a *complex polygenic disease*.

Let us assume, as before, that we have a response variable Y for a disease D and p predictors, $X_1, \ldots, X_p$, that contain the gene expression for genes $1, \ldots, p$.

4.3.1. Case 1

Let us assume disease D results from the simultaneous additive effect of several genes but there is only one unique way of obtaining D. In this case, the central subspace will also have the same dimension as in the mongenic case, since there is only one possible way of obtaining the disease. However, the difference arises from the form of the vector which constitutes a basis for the central subspace, b_1. As opposed to having only one non-zero b_{i1}, it will have several b_{i1}'s distinct from zero corresponding to the genes that have an effect. Let us consider an example to see this more clearly.

To keep things simple, let us assume that disease D is obtained by an additive effect of genes 1, 3 and 5. Then $b_1 = (b_{11}, 0, b_{31}, 0, b_{51}, 0, \ldots, 0)^T$ and the basis matrix B is $B = \Big[[b_1]\Big]$. Thus,

$$\begin{aligned}
B^T X &= \begin{bmatrix} b_{11} & 0 & b_{31} & 0 & b_{51} & \ldots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ \vdots \\ X_p \end{bmatrix} \\
&= b_{11} \times X_1 + 0 \times X_2 + b_{31} \times X_3 + \ldots + b_{51} \times X_5 + \ldots + 0 \times X_p \\
&= b_{11} X_1 + b_{31} X_3 + b_{51} X_5
\end{aligned}$$

which gives us the information required. So, if we determined the matrix B only containing a vector of the form of b_1, we would conclude that the disease has a central subspace of dimension 1 and arises from the simultaneous effect of the genes identified by the non-zero elements of b_1, where b_1 is a p-vector with at least two non-zero components.

4.3.2. Case 2

Let us now assume that the disease D can be obtained in d independent ways. For example, suppose that disease D can be obtained in four different ways:

(i) the over expression of gene 3 at some level
(ii) the over expression of gene 1 and under expression of gene 2
(iii) the simultaneous effect of gene 1, 2 and 4
(iv) the simultaneous effect of gene 1, 2 and 3
(v) mixture of all of the above

In this case we would have three basis vectors b_1, b_2 and b_3 where $b_1 = (0, 0, b_{13}, 0, \dots, 0)^T$, $b_2 = (b_{12}, b_{22}, 0, 0, \dots, 0)^T$ and $b_3 = (b_{13}, b_{23}, 0, b_{43} \dots, 0)^T$, for cases (i), (ii) and (iii). Notice that for (iv) and (v) we do not require another vector in the basis since these are included in span$\{b_1, b_2, b_3\}$. So, for example, case (iv) would be obtained from a linear combination of b_1, b_2 and b_3, that is,

$$\alpha b_1 + \beta b_2 + \gamma b_3, \quad \alpha, \beta, \gamma \in \mathbb{R}.$$

In particular, if we look more carefully we can see that (iv) is obtained solely from b_1 and b_2. So $\gamma = 0$ and thus we would have that case (iv) is obtained from

$$\alpha b_1 + \beta b_2 \quad \alpha, \beta \in \mathbb{R}.$$

The basis matrix B for the central subspace of disease D is given by

$$B = \begin{bmatrix} 0 & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ b_{31} & 0 & 0 \\ 0 & 0 & b_{43} \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix} \tag{3}$$

and for $B^T X$ we have

$$\begin{aligned} B^T X &= \begin{bmatrix} 0 & 0 & b_{31} & 0 & \dots & 0 \\ b_{12} & b_{22} & 0 & 0 & \dots & 0 \\ b_{13} & b_{23} & 0 & b_{43} & \dots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \vdots \\ X_p \end{bmatrix} \\ &= \begin{bmatrix} b_{31}X_3 \\ b_{12}X_1 + b_{22}X_2 \\ b_{13}X_1 + b_{23}X_2 + b_{43}X_4 \end{bmatrix} \\ &= \Big(b_{31}X_3, b_{12}X_1 + b_{22}X_2, b_{13}X_1 + b_{23}X_2 + b_{43}X_4 \Big)^T. \end{aligned}$$

Then a basis for the central subspace of D would be

$$\mathcal{B} = \{(0, 0, b_{13}, 0, \dots, 0), (b_{12}, b_{22}, 0, 0, \dots, 0), (b_{13}, b_{23}, 0, b_{43} \dots, 0)\}.$$

4.3.3. Central subspace for different forms of a disease

Let us now assume that we have a response variable Y for a disease but the response refers to different forms of the disease, say $D1$ and $D2$. In this case, central subspaces can also be used. However, now the different vectors $b_1, \ldots, b_d$ will give us information on different linear combinations responsible for obtaining different forms of disease or the different types of classifications of disease. For example, let us assume that the central subspace exists for the Golub et al. (1999) data – a well known Bioinformatics dataset and recently used in Han, Huang and Zhou (2021) – where the response variable denotes whether an individual has type ALL or type AML of leukaemia. And let us assume that the corresponding basis B for these two types of leukaemia is that given by (3). Then, we would say that the vectors

$$\mathcal{B} = \{(0, 0, b_{13}, 0, \ldots, 0), (b_{12}, b_{22}, 0, 0, \ldots, 0), (b_{13}, b_{23}, 0, b_{43} \ldots, 0)\}$$

span the central subspace for AML and ALL types of leukaemia.

4.4. Application to real datasets

In this section we apply three distinct methods to two different datasets, one where $n > p$ and another where $n \ll p$. For the case where $n > p$ we consider a classification problem and use the SIR method readily available in the R package dr (Weisberg, 2002) and the Lasso-SIR of Lin, Zhao and Liu (2019) readily available in the R package LassoSIR of Zhao, Lin and Liu (2017). We compare these two methods to a standard logistic regression. Although it is known that, in some instances, applying SIR based methods to classification problems can be problematic, in this instance, SIR performs extremely well. For further details on the problems that can arise when applying SIR to a classification setting see Cook and Yin (2001) and for recent developments see Zhang, Mai and Zou (2020). For the case where $n \ll p$, we consider a regression setting and use the SIR method available in the R package edrGrpahicalTools (Coudret, Liquet and Saracco, 2017), the LassoSIR (as in the $n < p$ case) and compare them to the standard Lasso (Tibshirani, 1996) available in the R package glmnet of Simon et al. (2011). As this is for illustration purposes, we do not go into details of estimating the dimension of the central subspace or estimating the number of slices required for the SIR method.

4.4.1. Case when $n > p$

We illustrate the use of central subspaces methods when $n > p$ with the well-known Breast Cancer Wisconsin (Diagnostic) Data Set, available at the Machine Learning repository `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`. The data set – first used in Street, Wolberg and Mangasarian (1993) and Wolberg, Street and Mangasarian (1994)

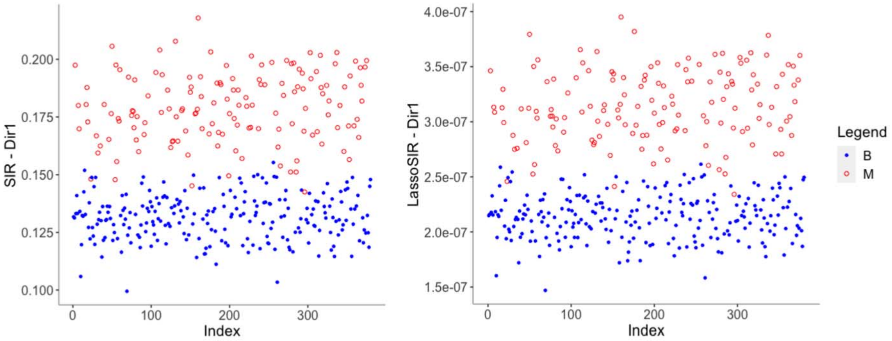

Fig 1. *Breast Cancer Wisconsin Data projected onto the basis vector of the central subspace estimated by SIR (left) and by Lasso-SIR (right).*

and more recently analysed in Ramsay, Durocher and Leblanc (2021) and Zhang and Lang (2022), for example – contains 569 samples of features on 30 variables. These are from digitized images of a fine needle aspirate of breast mass. There are 357 benign (B) samples and 212 malignant (M).

To estimate a basis for the central subspace of this problem – image representation of benign and malignant breast cancer – we first split the data into a training and a test set using a 70:30 split. We then apply SIR and the Lasso-SIR, with $h = 2$ slices and $d = 1$, the dimension of the central subspace†, to the training set to obtain an estimate of a basis vector for the central subspace. Once we have obtained the estimate of the basis vector for the central subspace we then transform the data by projecting it on to the central subspace basis vector and apply a logistic regression to perform prediction on the test set. Figure 1 displays the training data projected on to the central subspace basis vector estimated by SIR and by the Lasso-SIR, from which we can see that there is close to a linear separation between the classes malignant (M) and benign (B).

For comparison purposes we fit a logistic regression on the non-transformed training set and use this model to also make predictions on the test set. To compare the three methods, we use the ROC plot functionalities and comparison measures readily available in the R package hmeasure (Anagnostopoulos and Hand, 2012). These measures include the AUC, the H-measure (Hand, 2009), the Gini-index, sensitivity, specificity, recall and precision. We also include the number of variables used, in the creation of each model, *Num vars*. Note that for SIR and the Lasso-SIR this relates to the number of variables used to create the estimated basis vector of the central subspace, which in this case has dimension 1. Figure 2 and Table 3 summarise the results for the three methods on the test set showing an overall better performance for both the SIR+LR and the LassoSIR+LR methods in comparison to the standard logistic regression, LR.

From the last column in Table 3 – *Num vars* – we can see that the SIR method estimates a basis vector of the central subspace that is a linear combination of all the 30 variables available in the dataset. On the other hand, the Lasso-SIR

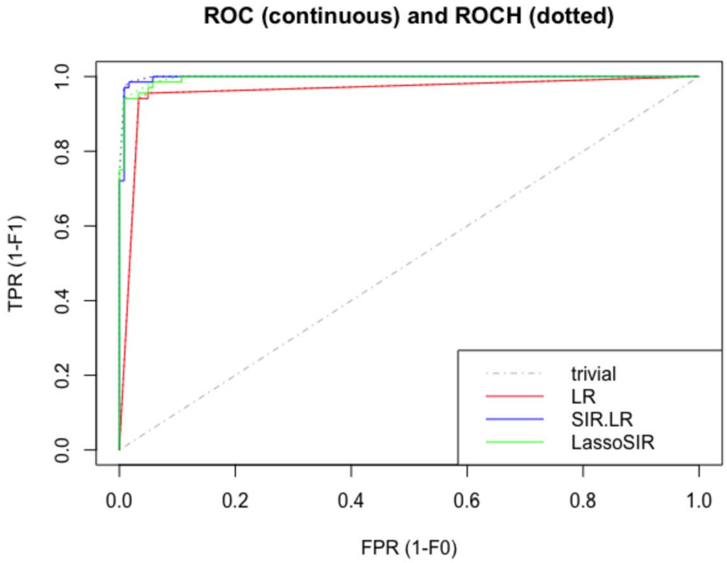

Fig 2. *ROC curves for the LR, SIR+LR and Lasso-SIR methods*

Table 3
Test set summary of comparative measures

Method	H	Gini	AUC	Sens	Spec	Precision	Recall	TPR	FPR	Num vars
LR	0.865	0.920	0.960	0.941	0.959	0.928	0.941	0.941	0.041	30
SIR+LR	0.956	0.994	0.997	0.926	0.992	0.984	0.926	0.926	0.008	30
Lasso-SIR+LR	0.920	0.990	0.995	0.941	0.992	0.985	0.941	0.941	0.008	19

method selects 19 variables out of the 30 and then creates a linear combination of these to estimate a basis vector for the central subspace. Therefore, although they both seem to estimate a basis vector for the central subspace, the Lasso-SIR estimate is more sparse. Using the Lasso-SIR estimate and the framework presented we would say that the central subspace of the image representation of benign and malignant breast cancer problem has dimension 1 and a basis vector for it is given by $b_1^T X$ where in this case b_1 has non-zero elements at positions $i \in \{2, 5, 6, 7, 8, 11, 14, 15, 17, 18, 19, 21, 22, 24, 25, 27, 28, 29, 30\}$. These correspond to the variables: *texture mean, smoothness mean, compactness mean, concavity mean, concave points mean, radius se, area se, smoothness se, concavity se, concave points se, symmetry se, radius worst, texture worst, area worst, smoothness worst, concavity worst, concave points worst, symmetry worst* and *fractal dimension worst*. The R code used for the analysis is available from the authors.

4.4.2. Case when $n \ll p$

To illustrate the application of these methods when $p > n$ we apply the SIR-QZ method of Coudret, Liquet and Saracco (2017) and the Lasso-SIR method (Lin, Zhao and Liu, 2019) to a subset of the Scheetz et al. (2006) dataset as done in Breheny and Huang (2013). The dataset consists of normalized microarray gene expression data harvested from the eye tissue of 120 twelve-week-old male rats. The outcome of interest is the expression of TRIM32, a gene which has been shown to cause Bardet-Biedl syndrome (Chiang et al., 2006) and the aim of the study was to determine genes associated with Bardet-Biedl syndrome. Here we use the data to estimate a basis for the central subspace that spans the space of those genes associated with the disease. The dataset contains approximately $p = 5000$ features (one-sixth of the original data) on $n = 120$ samples and was obtained by following the processing steps as detailed in Breheny and Huang (2013). Note that the original version of the SIR-QZ algorithm does not return the EDR directions. We changed the original code so that it returns these instead, so that we can project the data onto these directions, as was done in the previous example. As before, we also split the data so that 2/3 is used in the training set and 1/3 is used in the test set.

To apply SIR-QZ and the Lasso-SIR we first must specify the number of slices, h, used to calculate the edr directions and the dimension of the subspace, d. As h and d are often unknown, here, for exemplary purposes, we let $d = 1$ and use the training set to tune the h parameter. For both methods, we let h vary from 2 to 5 (as $h > d$ – this is a requirement of SIR-QZ) and for each value of h we: 1) determine the respective edr directions; 2) transform the data, by projecting it onto the respective edr directions; and, 3) fit a linear regression (LR) using the newly transformed data and calculate the mean-squared error (MSE) for the model. Note that the Lasso-SIR also requires tuning the regularization parameter λ, this is done at step 1) using the built-in cross-validation functionality and choosing the Lasso-SIR model associated with $\lambda = \mathsf{lambda.min}$, which aims at minimizing the cross-validation MSE over all 10-folds. Table 4 contains the results obtained on the training set for the SIR-QZ model and Lasso-SIR methods. Once the best model is obtained from the training set for each of the methods, it is then fit to the test set, on which both the SIR-QZ+LR and the Lasso-SIR + LR obtained an MSE of approximately 0.012. Once the best model is obtained from the training set for each of the methods, it is then fit to the test set, on which both the SIR-QZ+LR and the Lasso-SIR + LR obtained an MSE of approximately 0.012.

Table 4

Training set MSE and AIC values for each value of h for the SIR-QZ and Lasso-SIR methods.

Method	h	MSE	AIC	Num. vars
SIR-QZ	2	0.0141	−107.6733	5000
	3	0.0137	−110.1318	5000
	4	0.0119	−121.1168	5000
	5	0.0124	−117.9847	5000
Lasso-SIR	2	0.0111	−127.3761	7
	3	0.0106	−131.0188	12
	4	0.0094	−140.6779	6
	5	0.0106	−130.7025	3

For comparison purposes, we also fit a Lasso regression model to the data on the same training and test set using the R package glmnet (Simon et al., 2011). The lasso fitted model that is applied to the test set is obtained, as before, by choosing the model with $\lambda =$ lambda.min on the training set. The latter is defined by the authors as the model whose MSE, on the training data, with the minimum MSE. Note that these models are obtained on the training set using the default 10-fold cross validation. For further details see Friedman, Hastie and Tibshirani (2010). Results comparing the three models are given in Table 5.

Table 5
Results for the Lasso, SIR-QZ + linear regression and Lasso-SIR + linear regression models.

Model	Training MSE	Test MSE	Num. vars
Lasso	0.0169	0.0108	18
SIR-QZ, h = 4 + LR	0.0120	0.0125	5000
Lasso-SIR, h =4 + LR	0.0094	0.0120	6

From tables 4 and 5 we can see that in terms of MSE the SIR-QZ and the Lasso-SIR methods have similar performances, with the Lasso-SIR performing better on the training set in comparison to SIR-QZ. Both methods seem to agree on the number of slices h being 4. With respect to the number of variables used, as stated previously, SIR-QZ does not perform variable selection and uses all 5000 predictors to create an estimate of a basis vector, which is a linear combination of all the variables whereas the Lasso-SIR chooses 6 variables and estimates the basis vector using a linear combination of these 6 variables. Using the Lasso-SIR estimate and the framework presented we would say that the central subspace of the genes associated with TRIM32, and consequently the Bardet-Biedly syndrome, has dimension 1 and a basis vector for it is given by $b_1^T X$ where in this case b_1 has non-zero elements at positions $i \in \{2195, 3543, 3801, 4069, 4158, 4413\}$. These variables correspond to the the gene probes 1381902_at, 1390574_at, 1391916_at, 1393369_at, 1393743_at and 395332_at. The R code used for the analysis is available from the authors.

5. Summary and discussion

In this paper we have given a brief overview of the theory of central subspaces and reviewed some of the main methods that exist for determining the central subspace in practice. We have also developed a new classification system based on four criteria, namely, type of regression (inverse or non-inverse), type of subspace (linear or non-linear), dimensions of the data the method can be applied to ($p \leq n$ or $p > n$), and size of the data, regular or big. Using this classification we believe researchers can easily identify which central subspace method can be applied to their data and what type of central subspace they obtain once the method has been applied. It would also be useful to indicate whether the method can be applied to continuous and/or categorical data, however, we leave this to future research.

We also show how the the theory of central subspaces can be used in bioinformatics to classify different types of diseases according to the dimension of the central subspace and the form of the basis vectors. Note that the elements of B^TX, $b_1^TX, \dots, b_d^TX$, may or may not give a pathway to disease. The pathways may be complex functions arising from the central subspace. By determining B^TX we are simply determining the predictors (or combinations of predictors) that might be involved in the pathway. To see this more clearly, consider the complex polygenic disease, where we assumed that a basis for the central subspace of the disease would be

$$\mathcal{B} = \{(0, 0, b_{13}, 0, \dots, 0), (b_{12}, b_{22}, 0, 0, \dots, 0), (b_{13}, b_{23}, 0, b_{43} \dots, 0)\}$$

and

$$B^TX \quad = \quad (b_{31}X_3, b_{12}X_1 + b_{22}X_2, b_{13}X_1 + b_{23}X_2 + b_{43}X_4)^T$$

The function that we wish to estimate will take as arguments the elements of B^TX and possibly transform them into a more complex function. Examples of such functions are:

1. $f(B^TX) = b_{31}X_3 \times (b_{12}X_1 + b_{22}X_2) + \alpha\,(b_{13}X_1 + b_{23}X_2 + b_{43}X_4)$.
2. $f(B^TX) = \alpha_1(b_{31}X_3) + \alpha_2(b_{12}X_1 + b_{22}X_2) + \alpha_3(b_{13}X_1 + b_{23}X_2 + b_{43}X_4)$.
3. $f(B^TX) = (\alpha_1(b_{31}X_3) + \alpha_2(b_{12}X_1 + b_{22}X_2), \alpha_3(b_{13}X_1 + b_{23}X_2 + b_{43}X_4))$, note that here the response is multivariate.
4. $f(B^TX) = \mathbb{I}(X_3 > c_1) + \mathbb{I}(X_1 + X_2 > c_2) + \mathbb{I}(X_1 + X_2 + X_4 < c_3)$, where $c_1, c_2, c_3 \in \mathbb{R}$ and $\mathbb{I}$ denotes the indicator function.
5. $f(B^TX) = \alpha\mathbb{I}(X_1 + X_2 > c_1)\mathbb{I}(X_3 > c_2)\mathbb{I}(X_1 + X_2 + X_3 > c_3)$ for $\alpha, c_1, c_2, c_3 \in \mathbb{R}$.

However, we can only attempt to determine f after determining the matrix B and the elements of B^TX.

Another important aspect to consider is: does the central subspace always exist? As mentioned earlier, Cook (1998) gives conditions for its existence. In theory these conditions are not very restrictive but even if they hold we might not be able to determine a basis for the central subspace. To determine a basis for the central subspace for, say a gene expression disease, we need to guarantee that we have observed the expression on all elements that are relevant to the disease. Thus, if a gene expression study does not contain the expression of all relevant genes, when applying a central subspace technique to the data, it will be difficult to guarantee that the elements determined form a basis for the central subspace. This issue does give rise to a whole new set of questions: How does the DRS obtained in such a case compare to the actual central subspace? What happens if we have only observed elements that are correlated with the elements that are in the central subspace? and so on. We forsee that with the advancement of technology and the decreasing cost of sequencing technologies these questions will not be an issue. However, until then it would be useful to have answers to such questions. We leave these as open problems for futre research.

Acknowledgments

The authors sincerely thank the editor, the associate editor, and anonymous referees for their valuable comments that helped in improving the article significantly.

References

Adragni, K. and Raim, A. (2014). ldr: An R Software Package for Likelihood-Based Sufficient Dimension Reduction. *Journal of Statistical Software, Articles* **61** 1–21.

Anagnostopoulos, C. and Hand, D. J. (2012). hmeasure: The H-measure and other scalar classification performance metrics R package version 1.0.

Basu, D. and Pereira, C. A. B. (1983). Conditional Independence in Statistics. *Sankyā* **45** 324–337. MR0747602

Bernard-Michel, C., Gardes, L. and Girard, S. (2008). A Note on Sliced Inverse Regression with Regularizations. *Biometrics* **64** 982–984. MR2526650

Breheny, P. and Huang, J. (2013). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing* **25** 173–187. MR3306699

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351. MR2382644

Chen, C. H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8** 289–316. MR1624402

Chen, X. and Xie, M.-g. (2014). A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data. *Statistica Sinica* **24** 1655–1684. MR3308656

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J. et al. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet–Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences* **103** 6287–6292.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences* **176** 123–144. MR1869195

Cook, R. D. (1994a). On the interpretation of regression plots. *Journal of the American Statistical Association* **89** 177–189. MR1266295

Cook, R. D. (1994b). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section on Physical and Engineering Sciences* 18–25. American Statistical Association.

Cook, R. D. (1998). *Regression graphics. Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley and Sons Inc., New York. MR1645673

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics* **32** 1062–1092. MR2065198

Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science* **22** 1–26. MR2408655

Cook, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application* **5** 533–559. MR3774758

Cook, R. D. and Forzani, L. (2008a). Covariance reducing models: An alternative to spectral modelling of covariance matrices. *Biometrika* **95** 799–812. MR2461212

Cook, R. D. and Forzani, L. (2008b). Principal Fitted Components for Dimension Reduction in Regression. *Statistical Science* **23** 485–501. MR2530547

Cook, R. D. and Forzani, L. (2009). Likelihood-Based Sufficient Dimension Reduction. *Journal of the American Statistical Association* **104** 197–208. MR2504373

Cook, R. D., Forzani, L. and Tomassi, D. (2011). LDR: a package for likelihood-based sufficient dimension reduction. *Journal of statistical software* **39**. MR2643537

Cook, R. D., Forzani, L. and Rothman, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics* **40** 353–384. MR3014310

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94** 569–584. MR2410009

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to Achieve Elliptically Contoured Covariates in Regression. *Journal of the American Statistical Association* **89** 592–599.

Cook, R. D. and Ni, L. (2005). Sufficient Dimension Reduction via Inverse Regression. *Journal of the American Statistical Association* **100** 410–428. MR2160547

Cook, R. D. and Weisberg, S. (1991). Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association* **86** 328–332. MR1137117

Cook, R. D. and Weisberg, S. (1999). Graphs in statistical analyses: Is the medium the message? *The American Statistician* **53** 29–37.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis. *Australian & New Zealand Journal of Statistics* **43** 147–199. MR1839361

Coudret, R., Liquet, B. and Saracco, J. (2014). Comparison of sliced inverse regression approaches for underdetermined cases. *Journal de la Société Française de Statistique* **155** 72–96. MR3211755

Coudret, R., Liquet, B. and Saracco, J. (2017). edrGraphicalTools: Provides Tools for Dimension Reduction Methods R package version 2.2.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B. Methodological* **41** 1–31. MR0535541

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97** 279–294. MR2650738

Emerson, J. W. and Kane, M. J. (2012). Don't drown in the data. *Signifi-*

cance **9** 38–39.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research* **5** 73–99. MR2247974

Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics* **37** 1871–1905. MR2533474

Fukumizu, K. and Leng, C. (2014). Gradient-Based Kernel Dimension Reduction for Regression. *Journal of the American Statistical Association* **109** 359–370. MR3180569

Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica* **12** 1093–1113. MR1947065

Girard, S., Lorenzo, H. and Saracco, J. (2022). Advanced topics in Sliced Inverse Regression. *Journal of Multivariate Analysis* **188** 104852. MR4353861

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

Guha, S., Hafen, R., Rounds, J., Xia, J., Li, J. and Xi, B. (2012). Large complex data: divide and recombine (D&R) with RHIPE. *Stat* **1** 53–67. MR4027414

Hall, P. and Li, K. C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *The Annals of Statistics* **21** 867–889. MR1232523

Han, Y., Huang, L. and Zhou, F. (2021). A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. *Bioinformatics* **37** 2183–2189.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* **77** 103–123. MR3135860

Hang, W. and Xia, Y. (2017). MAVE: Methods for Dimension Reduction R package version 1.2.9.

Hilafu, H. and Yin, X. (2017). Sufficient Dimension Reduction and Variable Selection for Large- p-Small- nData With Highly Correlated Predictors. *Journal of Computational and Graphical Statistics* **26** 26–34. MR3610404

Kim, M. and Pavlovic, V. (2011). Central subspace dimensionality reduction using covariance operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33** 657–670.

Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics* **41** 221–249. MR3059416

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–342. MR1137117

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87** 1025–1039. MR1209564

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R. Monographs on Statistics and Applied Probability 161.* Taylor and Francis Group, LLC., New York. MR3838449

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics* **37** 1272–1298. MR2509074

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20** 3406–3412.

Li, L. and Nachtsheim, C. J. (2006). Sparse Sliced Inverse Regression. *Technometrics* **48** 503–510. MR2328619

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131. MR2422826

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics* **33** 1580–1616. MR2166556

Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and its Interface* **4** 73–83. MR2775250

Lin, Q., Zhao, Z. and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46** 580–610. MR3782378

Lin, Q., Zhao, Z. and Liu, J. S. (2019). Sparse Sliced Inverse Regression via Lasso. *Journal of the American Statistical Association* **114** 1726–1739. MR4047295

Liquet, B. and Saracco, J. (2011). A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approaches. *Computational Statistics* **27** 103–125. MR2877813

Liquet, B. and Saracco, J. (2016). BIG-SIR a sliced inverse regression approach for massive data. *Statistics and its Interface* **9** 509–520. MR3553378

Lvovs, D., Favorova, O. O. and Favorov, A. V. (2012). A polygenic approach to the study of polygenic diseases. *Acta Naturae* **4** 59–71.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107** 168–179. MR2949349

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review* **81** 134–150. MR3047506

Nghiem, L. H., Hui, F., Müller, S. and Welsh, A. (2023). Sparse Sliced Inverse Regression via Cholesky Matrix Penalization. *Statistica Sinica, to appear.*

Nilsson, J., Sha, F. and Jordan, M. I. (2007). Regression on manifolds using kernel dimension reduction. In *ICML '07 Proceedings of the 24th international conference on machine learning* 697–704. ACM, New York, NY, USA.

Prendergast, L. A. (2005). Influence functions for sliced inverse regression. *Scandinavian Journal of Statistics* **32** 385–404. MR2204626

Prendergast, L. A. (2007). Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. *Biometrika* **94** 585–601. MR2410010

Prendergast, L. A. and Smith, J. A. (2010). Influence functions for dimension reduction methods: An example influence study of principal hessian direction analysis. *Scandinavian Journal of Statistics* **37** 588–611. MR2779638

Qian, W., Ding, S. and Cook, R. D. (2019). Sparse Minimum Discrepancy Approach to Sufficient Dimension Reduction with Simultaneous Variable Selection in Ultrahigh Dimension. *Journal of the American Statistical Association* **114** 1277–1290. MR4011779

Ramsay, K., Durocher, S. and Leblanc, A. (2021). Robustness and asymptotics of the projection median. *Journal of Multivariate Analysis* **181** 104678. MR4157687

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C. and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America* **103** 14429–14434.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* **39** 1–13.

Street, W. N., Wolberg, W. H. and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis</title>. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology* (R. S. Acharya and D. B. Goldgof, eds.) 861–870. SPIE.

Tan, K., Shi, L. and Yu, Z. (2020). Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics* **48** 64–85. MR4065153

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological* **58** 267–288. MR1379242

Wang, H. and Xia, Y. (2008). Sliced Regression for Dimension Reduction. *Journal of the American Statistical Association* **103** 811–821. MR2524332

Weisberg, S. (2002). Dimension Reduction Regression in R. *Journal of Statistical Software, Articles* **7** 1–22.

Wolberg, W. H., Street, W. N. and Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters* **77** 163–171.

Wu, H.-M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics* **17** 590–610. MR2528238

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35** 2654–2690. MR2382662

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **64** 363–410. MR1924297

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 879–892. MR3382601

Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis* **101** 271–290. MR2557633

Zhang, S. and Lang, Z.-Q. (2022). Orthogonal least squares based fast feature selection for linear classification. *Pattern Recognition* **123** 108419.

Zhang, X., Mai, Q. and Zou, H. (2020). The Maximum Separation Subspace in Sufficient Dimension Reduction with Categorical Response. *Journal of Machine Learning Research* **21** 1-36. MR4073762

Zhao, Z., Lin, Q. and Liu, J. (2017). LassoSIR: Sparsed Sliced Inverse Regression via Lasso R package version 0.1.1. MR4047295

Zhong, W., Zeng, P., Ma, P., Liu, J. S. and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics* **21** 4169–4175.

Zhu, L.-X. and Fang, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **24** 1053–1068. MR1401836

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association* **101** 1638–1651. MR2279485

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67** 301–320. MR2137327