Graph informed sliced inverse regression

Eugen Pircalabelu[1], **Andreas Artemiou**[2]

[1] *UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences*
Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
eugen.pircalabelu@uclouvain.be;
[2] *Cardiff University, School of Mathematics*
Senghennydd Road 69, CF24 4AG Cardiff, Wales, UK
artemioua@cardiff.ac.uk;

Abstract

A new method is developed for performing sufficient dimension reduction when probabilistic graphical models are being used to perform estimation of parameters. The procedure enriches the domain of application of dimension reduction techniques to settings where (i) p the number of variables in the model is much larger than the available sample size n, (ii) p is much larger than the number of slices H the model uses and D the number of projection vectors can be larger than n. The methodology is developed for the case of the sliced inverse regression model, but extensions to other dimension reduction techniques such as sliced average variance estimation or other methods are straightforward.

Keywords: dimension reduction, sliced inverse regression, sliced average variance estimation, penalized estimation

1 Motivation

Sufficient dimension reduction (SDR) techniques and probabilistic graphical models (PGMs) have at their core the same objective, namely obtaining a map of the conditional independence statements that hold in the joint distribution of a multivariate vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\mathsf{T}$ with p components. The approaches that each methodological framework takes, are at first sight quite different from each other. The PGMs framework gives importance to the identification of (separator) nodes which, when conditioned upon, contain all the necessary information to make other groups of nodes conditionally independent of each other. The SDR framework gives importance to the identification of linear/nonlinear combinations of nodes which, when conditioned upon, retain all the necessary information to make groups of nodes conditionally independent of each other. In the SDR literature, the SIR model (Li, 1991) is one of the most popular and used techniques to achieve dimension reduction. Even though throughout the years many other methods have been introduced, SIR still remains one of the popular choices and a clear competitor to benchmark against.

One common aspect between PGMs and SIR is the fact that the routinely used algorithm for SIR uses information regarding inverse covariance matrices in order to standardize the original variables to alleviate the effects induced by different scales of measurement. These inverse covariance matrices can easily and accurately be obtained by using graph estimation techniques. As such, this manuscript exploits connections that enrich the framework of dimension reduction by making use of graphs. Our contribution is to introduce and study a new high-dimensional SDR method that allows for the cases where:

(i) p the number of variables in the model is larger than the available sample size n,

(ii) more than $H - 1$ directions are to be estimated and

(iii) can be used when the slicing variable is binary.

The method is based on graphical models and as such it also offers insight into the dependence structure that governs the vector $\boldsymbol{X}$. We propose as well a computational algorithm to facilitate its implementation in practice. The proposed method improves on limitations that popular SIR suffers from, while being easy to implement and general enough to be adapted to SAVE (Cook and Weisberg, 1991) and other methods than those described here.

The manuscript is structured as follows. In Section 2 we briefly present sufficient dimension reduction techniques and the probabilistic graphical modeling framework. In Section 3 the proposed method and estimation aspects are presented, while theoretical properties of the method are investigated in Section 4. A simulation study is presented in Section 5 and a real case analysis is presented in Section 6. We finish with a discussion on the method in Section 7.

2 Introduction to Sufficient dimension reduction and Probabilistic graphs

In this section we introduce briefly the two frameworks on which our procedure is based. We highlight the common points between dimension reduction methods and graphs, with the purpose of exploiting connections that allow one to make use of graphs when performing SDR.

2.1 Sufficient dimension reduction models

In a regression setting, one sets out to find a relationship between a response variable Y (without loss of generality we assume it to be univariate) and a p-dimensional vector $\boldsymbol{X}$. The larger p gets relative to the available sample size n, the more difficult it is for traditional methods to make meaningful statistical inference and the harder it is to visualize and interpret the results. Thus naturally, one might be interested in achieving dimension reduction either through feature selection (when a subset of the original variables is selected) or through feature extraction (when functions of the original predictors are extracted).

Sufficient dimension reduction is a class of methods proposed to achieve supervised feature extraction in a regression setting. It has been proposed for cases where one is interested in identifying D functions of the predictors without losing information for the conditional distribution of $Y|\boldsymbol{X}$. In other words, one is interested in estimating the $p \times D$ matrix $\boldsymbol{\beta}$ such that:

$$Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}. \tag{1}$$

This is known as the linear conditional independence model as the functions of the predictors one is extracting, are linear. The space spanned by the column vectors of $\boldsymbol{\beta}$ is known as the *Dimension reduction subspace (DRS)* which is denoted with $\mathcal{S}(\boldsymbol{\beta})$. There are multiple $\boldsymbol{\beta}$'s that satisfy model (1) and as such, the target is most often the space with the minimum dimension D, which is known as the *Central dimension reduction space* or simply the *Central space (CS)* and is denoted with $\mathcal{S}_{Y|\boldsymbol{X}}$. The CS does not always exist, but conditions of existence are mild and therefore we assume its existence throughout this work. On a positive note, if the CS exists it is unique and it can be shown to be the intersection of all DRS's.

SDR started with the breakthrough work of Li (1991) that introduced Sliced Inverse Regression (SIR) as a first method for SDR. A number of methods were proposed in the following years. A small sample includes Sliced Average Variance Estimate (SAVE) by Cook and Weisberg (1991), principal Hessian direction (pHd) by Li (1992), Minimum Average Variance Estimation (MAVE) by Xia et al. (2002), Contour Regression (CR) by Li et al. (2005), Directional Regression (DR) by Li and Wang (2007), Principal Support Vector Machines (PSVM) by Li et al. (2011) and many more. We refer to Ma and Zhu (2013) for a review on SDR and some of the proposed techniques.

Although SIR was the first method proposed in a long list of methodology, it is still the most popular method in the SDR literature and it is still used as a benchmark for any new method to compare to. This is due to its simplicity, since it first slices the response Y into H slices and then uses the matrix $\boldsymbol{\Lambda} = \text{cov}(\text{E}(\boldsymbol{X}|Y))$ to estimate the matrix $\boldsymbol{\beta}$. Note the use of the *inverse* expectation $\text{E}(\boldsymbol{X}|Y)$ rather than the usual $\text{E}(Y|\boldsymbol{X})$, hence the name.

SIR might be simple, but it is far from perfect. There are a number of shortcomings to it. First of all, it is routinely applied on standardized data. This implies that one has access to the inverse of the covariance matrix $\boldsymbol{\Sigma}_X = \text{cov}(\boldsymbol{X})$. For data where p is larger than n, the classical estimator for a covariance matrix used in SIR, is not per se invertible. Recently, Lin et al. (2019) proposed a SIR based procedure that avoids the need for an inverse covariance matrix.

The second problem is the dependence of the number of directions D that one can estimate, on the number of slices H. Due to the fact that the directions are extracted from $\boldsymbol{\Lambda}$, which is estimated using the predictor means in each slice, one is restricted in estimating at most $H-1$ directions. For example, in cases where Y is binary and one is restricted in taking only two slices, one can estimate at most one direction. This implies that one will not be able to extensively estimate the CS if it has a dimension larger than one and the response Y can be split only in two slices. This problem was recently tackled by Shin et al. (2017) where they extended the PSVM procedure of Li et al. (2011) by using weighted Support Vector Machines (SVM).

In recent years, sparse high-dimensional SDR methods have also been developed. We refer to the works of Li (2007), Wang and Yin (2008), Zhu and Zhu (2009), Radchenko (2015), Lin et al. (2018, 2019), Hilafu and Yin (2017) and Shin and Artemiou (2017) among others. What is common to all of the above works, is falling back on performing penalized regression models for estimation. The novelty of the method we propose in this manuscript, stems from exploring explicit connections that link probabilistic graphs and SDR due to their focus on conditional independence as was pointed out in Section 1.

In this paper, we address the above limitations and issues by proposing different estimators $\hat{\boldsymbol{\Sigma}}_X$ and $\hat{\boldsymbol{\Lambda}}$ that have their roots in the penalized *Gaussian graphical models* framework. Details on Gaussian graphical models are offered next, in Section 2.2.

2.2 Probabilistic graphical models

The first step in PGMs is to associate each component of the vector $(Y, X_1, \ldots, X_p)^\mathsf{T}$ with a node (and only one) in a graph $G(\mathcal{E}, \mathcal{V})$, where $\mathcal{E}$ denotes the set of undirected edges of the form $a-b$ between two nodes a and b that belong to the set of nodes $\mathcal{V} = \{1, \ldots, p+1\}$. We denote the edge set as $\mathcal{E} = \{(a,b)|\ a, b \in \mathcal{V}$ and a and b are connected by an edge in $G\}$.

This allows PGMs to posit conditional independence statements by identifying three disjoint sets of nodes $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ for which the relation $\mathcal{A} \perp\!\!\!\perp \mathcal{B}|\mathcal{C}$ holds. Conceptually, they translate the conditional independence statements that are encoded by a joint distribution function P, by identifying conditional independence statements that hold for nodes in the graph G. Switching between the graph statements and the distributional statements is allowed under a faithfulness assumption that links G and P. The assumption implies that all the conditional independence statements that can be read from the graph, hold as well for the joint distribution and vice-versa. This is denoted as

$$\mathcal{A} \perp\!\!\!\perp_G \mathcal{B}|\mathcal{C} \Leftrightarrow \mathcal{A} \perp\!\!\!\perp_P \mathcal{B}|\mathcal{C}. \tag{2}$$

For graphs, the *global* Markov property implies that for any disjoint node sets $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ from the graph G, for which $\mathcal{C}$ separates the nodes in the set $\mathcal{A}$ from the nodes in the set $\mathcal{B}$ (i.e. following any path formed by edges between a node in the set $\mathcal{A}$ and a node in the set $\mathcal{B}$ necessarily passes through a node in the separator set $\mathcal{C}$) the random variables included in the set $\mathcal{A}$ are conditionally independent of the variables in the set $\mathcal{B}$, given the random variables

included in the set $\mathcal{C}$. The toy example in Figure 1 illustrates the global Markov property that $Y \perp\!\!\!\perp X_2 \,|X_1$ since the path between $\mathcal{A} = \{Y\}$ and $\mathcal{B} = \{X_2\}$ passes through $\{X_1\}$, meaning that if X_1 and all edges connecting it to any other node are removed, then Y and X_2 would be completely disconnected.

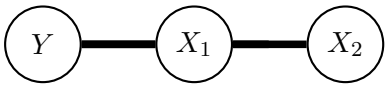

Figure 1: Toy example illustrating the global Markov property $Y \perp\!\!\!\perp X_2 \,|X_1$.

This property implies that identifying separator sets on graphs leads to conditional independencies that hold for the joint distribution. Moreover, the faithfulness assumption implies that these are *all* the conditional independence statements that hold for the joint distribution.

For the estimation of penalized sparse high-dimensional undirected graphs we refer to the works of Friedman et al. (2008), Ravikumar et al. (2008), Bickel and Levina (2008a,b), Boyd et al. (2011), Guo et al. (2011), Mazumder and Hastie (2012), Witten et al. (2011), Danaher et al. (2014) and Pircalabelu et al. (2016) among many others.

Under a Gaussian assumption for the vector $(Y, X_1, \ldots, X_p)^\mathsf{T}$, (2) is equivalent to (3) below:

$$(Y, X_1, \ldots, X_p)^\mathsf{T} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_Y & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_X \end{bmatrix}\right) \text{ and let } \boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} \text{ then } (a,b) \in \mathcal{E} \Leftrightarrow \boldsymbol{\Theta}_{ab} \neq 0. \tag{3}$$

Due to properties of the multivariate Gaussian distribution, the marginal model also holds:

$$\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X) \text{ and let } \boldsymbol{\Theta}_X = \boldsymbol{\Sigma}_X^{-1} \text{ then } (a,b) \in \mathcal{E}_X \Leftrightarrow \boldsymbol{\Theta}_{X;ab} \neq 0. \tag{4}$$

Moreover, conditionally on $\boldsymbol{X}$, Y also follows a Gaussian distribution:

$$Y|\boldsymbol{X} \sim N(\boldsymbol{\Sigma}_{YX}\boldsymbol{\Theta}_X\boldsymbol{X}, \sigma_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Theta}_X\boldsymbol{\Sigma}_{XY}). \tag{5}$$

3 Graph informed procedure

The conditional model in (5) is from a distributional point of view attractive, however it is relatively restrictive, since it specifies that $\mathrm{E}(Y|\boldsymbol{X})$ depends linearly on $\boldsymbol{X}$. Our method starts from the multiple index model commonly used in the SDR literature of the form

$$Y = g(\boldsymbol{\beta}_1^\mathsf{T}\boldsymbol{X}, \boldsymbol{\beta}_2^\mathsf{T}\boldsymbol{X}, \ldots, \boldsymbol{\beta}_D^\mathsf{T}\boldsymbol{X}, \epsilon) \quad \text{where} \tag{6}$$

(i) $\boldsymbol{X} \in \mathbb{R}^p$ with $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_X)$

(ii) $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D \in \mathbb{R}^p$

(iii) $\epsilon \perp\!\!\!\perp (X_1, \ldots, X_p)^\mathsf{T}$

(iv) $\mathrm{E}(\epsilon) = 0$

(v) $g : \mathbb{R}^{D+1} \to \mathbb{R}$.

The link function $g(\cdot)$ is an unknown many-to-one function, the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D$ are unknown vectors of coefficients and for simplicity we assume throughout the manuscript that the dimension D is fixed and known upfront. A number of proposals exist in the literature to estimate D if it is unknown, see for example sequential tests in Bura and Yang (2011), a BIC-type strategy like in Zhu et al. (2006) or the ladle plot idea in Luo and Li (2016).

Let $\boldsymbol{W}$ denote the random vector $\mathrm{E}(\boldsymbol{X}|Y)$ with covariance matrix $\boldsymbol{\Lambda} = \mathrm{cov}(\boldsymbol{W})$ for which $\mathcal{S}_{\boldsymbol{\Lambda}}$ is the space spanned by its columns. Let next $\boldsymbol{\beta}$ be the matrix of coefficients of dimension $p \times D$ obtained by staking the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D$ together. As pointed out in Section 2.1, under the assumptions specified in (6), $\boldsymbol{\beta}$ is not uniquely defined. However, the space $\mathcal{S}_{Y|\boldsymbol{X}}$ defined by the columns of $\boldsymbol{\beta}$ is uniquely defined. As $\boldsymbol{X}$ follows an elliptical distribution, one also has as a consequence that

$$\boldsymbol{\Sigma}_X \mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}_{\boldsymbol{\Lambda}}.$$

The goal is to recover consistently $\mathcal{S}_{Y|\boldsymbol{X}} = \boldsymbol{\Theta}_X \mathcal{S}_{\boldsymbol{\Lambda}}$ using sample information when n, the number of sample points, is smaller than p, the number of components and $\boldsymbol{\Theta}_X$ is sparse and graph structured.

Assume one has at disposal n i.i.d samples from the vector $(Y, \boldsymbol{X}^\mathsf{T})^\mathsf{T}$ denoted as $(Y_i, \boldsymbol{X}_i^\mathsf{T})^\mathsf{T}$ where $i = 1, \ldots, n$. In classical low-dimensional settings where $p < n$, one can consistently estimate $\boldsymbol{\Sigma}_X$ using the sample covariance matrix $\boldsymbol{S}_X = (1/n) \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T}$ (see Lin et al., 2018), but this no longer holds for high-dimensional settings. As such, under the assumption of normality, we estimate directly $\boldsymbol{\Theta}$ by using the graphical lasso procedure as it optimizes the penalized negative Gaussian log-likelihood function. The optimization routine takes the form:

$$\min_{\boldsymbol{\Theta}_X} \left\{ \mathrm{trace}(\boldsymbol{S}_X \boldsymbol{\Theta}_X) - \log\det \boldsymbol{\Theta}_X + \lambda_{n1} \sum_{a \neq b} \mid \boldsymbol{\Theta}_{X;ab} \mid \right\}, \tag{7}$$

such that $\boldsymbol{\Theta}_X \succ 0$ (positive definite) and where λ_{n1} is a known, positive regularization level that is specified by the user. The role of λ_{n1} is to determine the sparsity of $\boldsymbol{\Theta}_X$ as it shrinks small entries of $\boldsymbol{\Theta}_X$ to 0 values. The larger λ_{n1} is, the more sparsity is enforced in the model, as more entries get shrunk to 0. In general, the regularization parameter is not known upfront and a search on a grid is usually performed where the quality of the selected model is assessed either by information criteria or cross-validation. An optimizer to (7) has been proposed in Friedman et al. (2008) and can be obtained by using the *glasso* algorithm. Due to the relation in (4) we get with no extra effort an estimated undirected graph that summarizes all the conditional independence statements that hold for the components of $\boldsymbol{X}$.

We note that this allows to bypass the $n > p$ requirement, as now p can be much larger than n and still allow for the estimated covariance matrix to be positive definite. The high-dimensional sparse graph will provide the researcher extra information regarding the conditional independencies for the predictors. However attractive this property might be, it is not sufficient, since the number of directions one can extract is still bounded by $H - 1$, where H is the number of slices. In Sections 3.1 and 3.2 we will propose as well an ℓ_1 graph based strategy to overcome this limitation.

Recently, Lin et al. (2019) proposed a penalized ℓ_1 framework that is able to consistently recover the CS even when $p > n$. Their procedure does not tackle the estimation of the inverse covariance matrix $\boldsymbol{\Theta}$, but rather takes a nodewise approach (see Meinshausen and Bühlmann, 2006) where they fit a penalized model for each vector $\boldsymbol{\beta}_d$ with $d = 1, \ldots, D$. If D is small, then such a nodewise strategy, might result in considerable speed gains. We note that Lin et al. (2019) assume that each vector of coefficients is sparse, whereas our method assumes that the matrices $\boldsymbol{\Theta}_X$ and $\boldsymbol{\Omega}$ from equation 8 in Section 3.1 are both sparse and graph structured.

We illustrate next our proposed estimation method which we denote throughout the manuscript as 'GraphSIR', short for 'graph informed slice inverse regression'. As we have mentioned in Section 1, similar ideas can be used for other SDR techniques as well and to illustrate this we provide also a graph based extension for the SAVE method denoted as GraphSAVE.

3.1 The GraphSIR case

We present in this section the GraphSIR procedure as an extension of the SIR method to the high-dimensional case. The procedure follows a similar algorithmic description as for the classical low-dimensional case, but the extensions proposed in this section allow for the estimation of the effective dimension reduction (edr) directions when (i) $p > n$ and (ii) when the number of directions is larger than $H - 1$. We stress that under the classical SIR methodology settings (i)-(ii) cannot properly be dealt with. We overcome such limitations by deploying techniques from the PGM framework, thus enriching the SDR methodology.

Step 1: Solve (7) with the graphical lasso procedure and obtain the estimated value for $\hat{\boldsymbol{\Theta}}_X$.

Step 2: Using the estimate $\hat{\boldsymbol{\Theta}}_X$ (which is positive definite even if $p > n$) calculate $\hat{\boldsymbol{\Sigma}}_X^{-1/2}$.

Step 3: Standardize $\boldsymbol{X}_i$ by an affine transformation to get $\hat{\boldsymbol{Z}}_i = \hat{\boldsymbol{\Sigma}}_X^{-1/2}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$.

Step 4: Divide the range of Y_i into H non-overlapping slices and calculate

(i) the 'sample size' of the slice as: $n_h = \sum_{i=1}^{n} \delta_h(Y_i)$ where $\delta_h(Y_i)$ is the indicator function taking the values 0 or 1 depending on whether the value of Y_i falls in the h-th slice or not;

(ii) the empirical mean of each slice as: $\hat{\boldsymbol{m}}_h = (1/n_h) \sum_{i=1}^{n} \hat{\boldsymbol{Z}}_i \delta_h(Y_i)$.

Step 5: Under the assumption $\hat{\boldsymbol{m}}_h \approx N(0, \boldsymbol{\Lambda}) \quad \forall h = 1, \ldots, H$, where $\boldsymbol{Z} = \boldsymbol{\Sigma}_X^{-1/2}(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))$, optimize:

$$\min_{\boldsymbol{\Omega}} \left\{ \operatorname{trace}(\boldsymbol{S}_Z \boldsymbol{\Omega}) - \log\det \boldsymbol{\Omega} + \lambda_{n2} \sum_{a \neq b} \mid \boldsymbol{\Omega}_{ab} \mid \right\}, \tag{8}$$

such that $\boldsymbol{\Omega} \succ 0$. We denote by $\hat{\boldsymbol{\Omega}}$ the estimator obtained when optimizing (8) and the matrix $\boldsymbol{S}_Z = (1/n) \sum_{h=1}^{H} n_h \hat{\boldsymbol{m}}_h \hat{\boldsymbol{m}}_h^{\mathsf{T}}$ denotes the empirical covariance matrix based on the vectors $\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_H$.

Step 6: Perform an eigenvalue decomposition on the estimated covariance matrix $\hat{\boldsymbol{\Lambda}} \equiv \hat{\boldsymbol{\Omega}}^{-1}$ (which is positive definite even if $p > H$) and retain the eigenvectors $\hat{\boldsymbol{\eta}}_d$ where $d = 1, \ldots, D$ corresponding to the largest D eigenvalues. Transform the standardized edr directions $\hat{\boldsymbol{\eta}}_d$ back to the original scale by returning the estimators $\hat{\boldsymbol{\beta}}_d = \hat{\boldsymbol{\Sigma}}_X^{-1/2} \hat{\boldsymbol{\eta}}_d$ for $d = 1, \ldots, D$.

Note that λ_{n2} in Step 5 is a known, positive regularization level specified by the user, which dictates the sparsity level of the $\boldsymbol{\Omega}$ matrix. In practice a grid search is performed for an optimal value. Regardless of how large p is compared to n or H, the estimator $\hat{\boldsymbol{\Omega}}$ is always positive definite.

Step 6 overcomes a major shortcoming of the classical SIR method where the number of non-zero eigenvalues *could not* be larger than $H - 1$, which is problematic in case Y is binary. Our procedure ensures that $\hat{\boldsymbol{\Omega}}$ is full rank and thus the number of non-zero eigenvalues does not depend on H, implying that one can identify more directions than slices, which is not possible in the classical SIR approach. Moreover, Step 5 explicitly ensures positive definiteness which was not guaranteed in the classical SIR as there, an eigenvalue decomposition is performed directly on $\boldsymbol{S}_Z$, however its positive definiteness cannot be guaranteed. It is in Steps 5 and 6 that the novelty of the procedure and usefulness of the graph methods is manifested, however we acknowledge that since in Step 6 one uses the covariance matrix obtained as $\hat{\boldsymbol{\Omega}}^{-1}$, one might envision different strategies using other high-dimensional covariance matrix estimators.

3.2 The GraphSAVE case

We present in this section the GraphSAVE procedure as an extension of the SAVE method to the high-dimensional case. Since the procedural recipe for SAVE is almost identical to that of SIR, the only difference appears in Steps 5-6, where instead of $\boldsymbol{S}_Z$, one uses $\tilde{\boldsymbol{S}}_Z$ defined as the sliced average variance estimate of Cook and Weisberg (1991):

$$\tilde{\boldsymbol{S}}_Z = \sum_{h=1}^{H} (\boldsymbol{I} - \mathrm{cov}(\boldsymbol{Z}|Y \in \text{ slice } h))(\boldsymbol{I} - \mathrm{cov}(\boldsymbol{Z}|Y \in \text{ slice } h))^{\mathsf{T}},$$

with $\boldsymbol{Z}$ defined as previously and $\boldsymbol{I}$ being the identity matrix.

Steps 1-4: The same as for GraphSIR;

Step 5: Calculate in each slice the empirical covariance matrix of the $\hat{\boldsymbol{Z}}_i$'s as

$$\tilde{\boldsymbol{S}}_h = (1/n_h) \sum_{i=1}^{n} (\hat{\boldsymbol{Z}}_i \delta_h(Y_i) - \hat{\boldsymbol{m}}_h)(\hat{\boldsymbol{Z}}_i \delta_h(Y_i) - \hat{\boldsymbol{m}}_h)^{\mathsf{T}},$$

and minimize in each slice:

$$\min_{\boldsymbol{\Omega}_h} \left\{ \mathrm{trace}(\tilde{\boldsymbol{S}}_h \boldsymbol{\Omega}_h) - \log\det \boldsymbol{\Omega}_h + \lambda_{n2} \sum_{a \neq b} | \, \boldsymbol{\Omega}_{h;ab} \, | \right\},$$

such that $\boldsymbol{\Omega}_h \succ 0$.

Step 6: Calculate the estimated pooled covariance matrix over the slices as:

$$\hat{\boldsymbol{\Lambda}}_{\text{pooled}} = (1/n) \sum_{h=1}^{H} n_h (\hat{\boldsymbol{\Omega}}_h^{-1} - \boldsymbol{I})(\hat{\boldsymbol{\Omega}}_h^{-1} - \boldsymbol{I})^{\mathsf{T}}.$$

Perform an eigenvalue decomposition on the estimated covariance matrix $\hat{\boldsymbol{\Lambda}}_{\text{pooled}}$ and retain the eigenvectors $\hat{\boldsymbol{\eta}}_d$ where $d = 1, \ldots, D$ corresponding to the largest D eigenvalues. Transform the standardized edr directions $\hat{\boldsymbol{\eta}}_d$ back to the original scale by returning the estimators $\hat{\boldsymbol{\beta}}_d = \hat{\boldsymbol{\Sigma}}_X^{-1/2} \hat{\boldsymbol{\eta}}_d$ for $d = 1, \ldots, D$.

Note that for the GraphSAVE case, the procedure is computationally more intensive than for GraphSIR especially for a larger number of slices, since an ℓ_1 minimization is performed for each slice in Step 5. The consequence however is that, $\hat{\boldsymbol{\Lambda}}_{\text{pooled}}$ is always positive definite since it involves a summation of positive definite matrices.

4 Theoretical properties

We allow the number of nodes, the sparsity of the graph and the number of slices to depend on the sample size, and denote this by p_n, s_n and H_n.

Regularity conditions.

(a) there exist constants τ_1 and τ_2 such that

$$0 < \tau_1 \leq \mathrm{eig}_{\min}(\boldsymbol{\Sigma}_{X;0}) \leq \mathrm{eig}_{\max}(\boldsymbol{\Sigma}_{X;0}) \leq \tau_2 < \infty;$$

(b) $\lambda_{n1} = O\Big((1 + p_n/(s_{n1}+1))\sqrt{\log p_n/n}\Big)$, where

$$s_{n1} = \mid S_1 \mid -p_n$$
$$S_1 = \{(a,b) | \boldsymbol{\Theta}_{X;0;ab} \neq 0\},$$

with $\boldsymbol{\Theta}_{X;0;ab}$ denoting the element on position (a,b) from the true concentration matrix $\boldsymbol{\Theta}_{X;0} \equiv \boldsymbol{\Sigma}_{X;0}^{-1}$ and $\mid S_1 \mid$ denoting the cardinality of the set.

(c) there exist constants τ_3 and τ_4 such that

$$0 < \tau_3 \leq \text{eig}_{\min}(\boldsymbol{\Lambda}_0) \leq \text{eig}_{\max}(\boldsymbol{\Lambda}_0) \leq \tau_4 < \infty;$$

(d) $\lambda_{n2} = O\Big((1 + p_n/(s_{n2}+1))\sqrt{\log p_n/H_n}\Big)$, where

$$s_{n2} = \mid S_2 \mid -p_n$$
$$S_2 = \{(c,d) | \boldsymbol{\Omega}_{0;cd} \neq 0\},$$

with $\boldsymbol{\Omega}_{0;cd}$ denoting the element on position (c,d) from the true concentration matrix $\boldsymbol{\Omega}_0 \equiv \boldsymbol{\Lambda}_0^{-1}$ and $\mid S_2 \mid$ denoting the cardinality of the set.

Proposition 1. *Under regularity conditions (a)–(d) if* $\sqrt{\log p_n/n} = O(\lambda_{n1})$ *and* $\sqrt{\log p_n/H_n} = O(\lambda_{n2})$*, then there exist global minimizers of (7) and (8) that satisfy*

$$|| \hat{\boldsymbol{\Theta}}_X - \boldsymbol{\Theta}_{X;0} ||_F^2 = O_p((p_n + s_{n1}) \log p_n/n)$$
$$|| \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0 ||_F^2 = O_p((p_n + s_{n2}) \log p_n/H_n).$$

The proof comes as a direct application of Theorem 1 from Lam and Fan (2009) and is omitted. Under extra mild conditions as in Theorem 2 from Lam and Fan (2009), the graphs can be shown to be also sparsistent, meaning that the estimated '0' values in the matrices $\hat{\boldsymbol{\Theta}}_X$ and $\hat{\boldsymbol{\Omega}}$ are identified with high probability on the correct positions.

Proposition 2. *Under regularity conditions (a)–(d), we have that*

$$|| \hat{\boldsymbol{\Theta}}_X \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Theta}_{X;0} \boldsymbol{\Lambda}_0 ||_2 = O_p((\log p_n((p_n + s_{n1})/n + (p_n + s_{n2})/H_n))^{1/2}).$$

Proof. Using classical matrix norm properties we have that

$$\begin{aligned}
|| \hat{\boldsymbol{\Theta}}_X \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Theta}_{X;0} \boldsymbol{\Lambda}_0 ||_2 =& || \hat{\boldsymbol{\Theta}}_X \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Theta}_{X;0} \hat{\boldsymbol{\Lambda}} + \boldsymbol{\Theta}_{X;0} \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Theta}_{X;0} \boldsymbol{\Lambda}_0 ||_2 \\
=& || (\hat{\boldsymbol{\Theta}}_X - \boldsymbol{\Theta}_{X;0}) \hat{\boldsymbol{\Lambda}} + \boldsymbol{\Theta}_{X;0} (\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0) ||_2 \\
\leq& || (\hat{\boldsymbol{\Theta}}_X - \boldsymbol{\Theta}_{X;0}) \hat{\boldsymbol{\Lambda}} ||_2 + || \boldsymbol{\Theta}_{X;0} (\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0) ||_2 \\
\leq& || \hat{\boldsymbol{\Theta}}_X - \boldsymbol{\Theta}_{X;0} ||_2 || \hat{\boldsymbol{\Lambda}} ||_2 + || \boldsymbol{\Theta}_{X;0} ||_2 || \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0 ||_2 \\
\leq& || \hat{\boldsymbol{\Theta}}_X - \boldsymbol{\Theta}_{X;0} ||_F || \hat{\boldsymbol{\Lambda}} ||_2 + || \boldsymbol{\Theta}_{X;0} ||_2 || \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0 ||_2
\end{aligned}$$

where $|| \cdot ||_2$ and $|| \cdot ||_F$ denote the operator and Frobenius norms.

Now, due to symmetry and using the property of eigenvalues of matrix powers, we can rewrite

$$\begin{aligned}
|| \hat{\boldsymbol{\Lambda}} ||_2 &= \text{eig}_{\min}^{-1}(\hat{\boldsymbol{\Omega}}) \\
&= \text{eig}_{\min}^{-1}(\hat{\boldsymbol{\Omega}} + \boldsymbol{\Omega}_0 - \boldsymbol{\Omega}_0)
\end{aligned}$$

$$
\begin{aligned}
&\leq (\text{eig}_{\min}(\boldsymbol{\Omega}_0) + \text{eig}_{\min}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0))^{-1} \\
&= (O(1) + o_p(1))^{-1} \\
&= O_p(1).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
|| \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}_0 ||_2 =& || \hat{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0)\boldsymbol{\Lambda}_0 ||_2 \\
\leq& || \hat{\boldsymbol{\Lambda}} ||_2 || \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0 ||_2 || \boldsymbol{\Lambda}_0 ||_2 \\
\leq& || \hat{\boldsymbol{\Lambda}} ||_2 || \hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0 ||_F || \boldsymbol{\Lambda}_0 ||_2 \\
=& O_p(1) \times O_p(\sqrt{(p_n + s_{n2}) \log p_n / H_n}) \times O(1) \\
=& O_p(\sqrt{(p_n + s_{n2}) \log p_n / H_n}).
\end{aligned}
$$

As $|| \boldsymbol{\Theta}_{X;0} ||_2 = O(1)$, the claim follows directly after substitution. □

Proposition 2 can be seen as an analog of Theorems 2 and 6 from Lin et al. (2018) for our procedure. The point of departure from their results is the fact that we (i) drop the assumption $\lim p_n/n = 0$ as $n \to \infty$, (ii) use $\hat{\boldsymbol{\Theta}}_X$ as estimator, rather than $(1/n)\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T}$ or the thresholded covariance estimator of Bickel and Levina (2008a) and (iii) provide explicitly the convergence rate.

5 Simulation study

The performance of the GraphSIR and GraphSAVE procedures has been investigated in a controlled simulation study. The performance of each method was measured by

$$
\text{Loss} = || \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^\mathsf{T}\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^\mathsf{T} - \boldsymbol{\beta}_0(\boldsymbol{\beta}_0^\mathsf{T}\boldsymbol{\beta}_0)^{-1}\boldsymbol{\beta}_0^\mathsf{T} ||_F,
$$

where $\hat{\boldsymbol{\beta}}$ represents the estimated coefficient matrix, $\boldsymbol{\beta}_0$ represents the true coefficient matrix and $|| \cdot ||_F$ is the Frobenius norm. Lower values of the Loss indicate better performance.

We simulated data from the following models:

Model 1: $Y = X_1 + X_2 + \sigma\epsilon$ where $D = 1$;

Model 2: $Y = X_1/(.5 + (X_2 + 1)^2) + \sigma\epsilon$ where $D = 2$;

Model 3: $Y = X_1^2 + X_2 + \sigma\epsilon$ where $D = 2$;

Model 4: $Y = (\sqrt{X_1^2 + X_2^2}) \log(\sqrt{X_1^2 + X_2^2}) + \sigma\epsilon$ where $D = 2$;

Model 5: $Y = \sin(X_1) + \cos(X_2) + X_3 + X_4^2 + \sigma\epsilon$ where $D = 4$;

Model 6: $Y = X_1/X_2 + X_3 + \sigma\epsilon$ where $D = 3$.

In each model $\boldsymbol{X} = (X_1, \ldots, X_p)^\mathsf{T} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$ and $\rho \in \{.5, .9\}$. The errors ϵ followed $\epsilon \sim N(0, 1)$ and $\sigma \in \{.2, .5\}$. From each model we sampled $n = 100$ iid observations and for SIR-based methods the number of variables was set to $p \in \{10, 50, 100, 500\}$, while the number of slices was set to $H \in \{2, 5, 10, 20\}$. Due to computational complexity, for SAVE-based methods the number of variables was set to $p \in \{10, 50, 100\}$.

We ran GraphSIR, SIR and LassoSIR (Lin et al., 2019) for models 1, 2, 5 and 6, while GraphSAVE and SAVE are run using models 3, 4, 5 and 6. For models 1, 2, 3 and 4 only the first two variables are active, whereas the remaining $p - 2$ components of $\boldsymbol{X}$ have no influence on Y. For models 5 and 6, the first four and three variables are active, respectively.

For each scenario 100 independent repetitions from the same generating process were performed and all competitors know the true value of D. For GraphSIR and GraphSAVE the tuning parameters have been selected with the extended BIC criterion (Foygel and Drton, 2010) on a grid of 200 equally spaced values with hyper-parameter τ fixed at the value $= .5$. The criterion is defined as:

$$\text{BIC}_\gamma = -2\ell_n(\hat{\mathbf{\Psi}}(E)) + |E| \log n + 4|E|\tau \log p,$$

where $\ell_n(\cdot)$ is the maximized Gaussian log-likelihood obtained using the estimated concentration matrix generically denoted as $\hat{\mathbf{\Psi}}(E)$, $|E|$ is the cardinality of the estimated set of edges in the corresponding graph associated with $\hat{\mathbf{\Psi}}$, τ in a penalty factor and p and n denote the number of variables and sample size used in the estimation process. Note that when performing Step 1 of the GraphSIR algorithm $\hat{\mathbf{\Psi}} \equiv \hat{\mathbf{\Theta}}$, while in Step 5 $\hat{\mathbf{\Psi}} \equiv \hat{\mathbf{\Omega}}$. Similarly, in Step 5 of the GraphSAVE algorithm $\hat{\mathbf{\Psi}} \equiv \hat{\mathbf{\Omega}}_h$ for each slice $h = 1, \ldots H$.

For fitting LassoSIR models we have used the available `R`-based package `LassoSIR` accompanying the paper of Lin et al. (2019). The default k-fold cross-validation procedure was used to select the tuning parameters in this case. In the case of the penalized LassoSIR procedure, for some settings the procedure provides error messages and does not converge for any of the 100 repetitions and when LassoSIR did not converge, we have set its Loss index to 'NA'. Neither convergence issues nor error messages have been observed for GraphSIR and GraphSAVE.

Figure 2 presents the obtained results. In the plots each symbol represents the average loss over the 100 repetitions for each scenario. For the high-dimensional scenarios where $n < p$, the classical SIR and SAVE cannot be used and as such we set also their performance index to 'NA' for those instances.

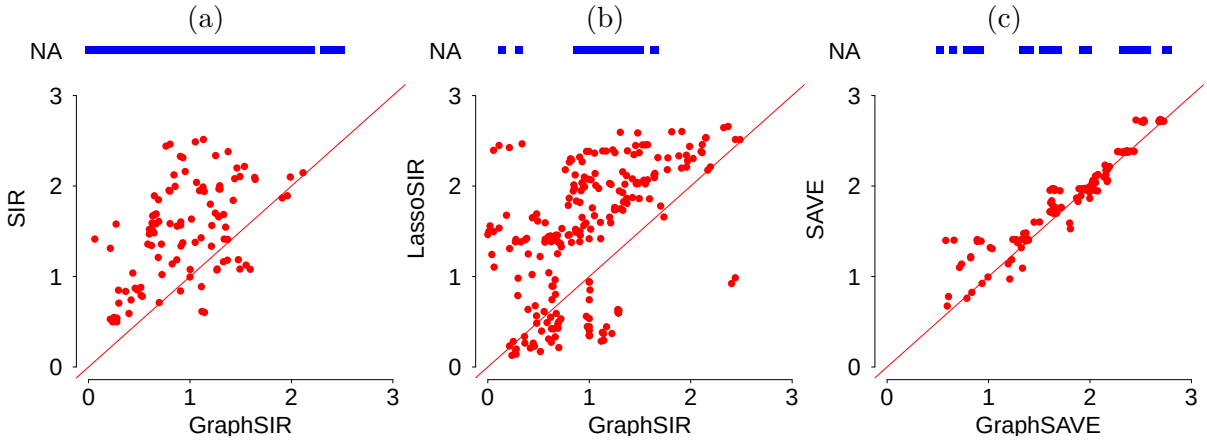

Figure 2: Simulation data. Frobenius loss (smaller is better) between $\hat{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_0$ for GraphSIR and GraphSAVE (x-axis) and competitors (y-axis). Values above the main diagonal show a better performance of GraphSIR and GraphSAVE compared to the competitors. Each symbol represents the empirical average of 100 repetitions from a scenario. The square symbols represent settings where SIR, LassoSIR or SAVE did not produce a numerical solution. The results of 256 different scenarios are plotted in panels (a) and (b) and the results of 192 different scenarios are plotted in panel (c).

When compared to the classical SIR and the LassoSIR method, see panels (a) and (b), the GraphSIR provided similar or better results in a large majority of different scenarios, even in the low-dimensional settings where $n > p$. When compared to the classical SAVE procedure in panel (c), the GraphSAVE method provided as well very similar results in all cases.

We zoom-in on the results obtained and present in Figures 3 and 4 the performance of the methods for the different values of the simulation parameters. Based on Figure 3 we conclude that increasing p is more detrimental for SIR than for GraphSIR, while at the same time the highly non-linear model 5 is more challenging for SIR than for GraphSIR. Figure 4 indicates that relative to the performance of GraphSIR, models 5 and 6 are challenging for LassoSIR, but for model 1 its performance seems to be better than that of GraphSIR. Moreover, for the cases where LassoSIR did not converge, the number of slices was always low, $H = 2$, pointing to the same limitation as for classical SIR of the dependence of the number of estimated directions on the number of slices, which is problematic when H is small and $D \geq H$.

We conclude that compared to the low-dimensional methods, the newly proposed graph informed methods provide similar or better results for cases when $n > p$, with the added value that they can also handle high-dimensional cases. Compared to the high-dimensional LassoSIR method, the graph informed SIR proved to be a worthy competitor as it provided comparable and close results, while in many settings being more stable in terms of convergence.

6 Real case

In this section we analyze a simplified version of the gene expression data from Scheetz et al. (2006) to illustrate the performance of the GraphSIR and GraphSAVE methods. The data contain information about the expression level of $p = 200$ genes (predictors) for a total of $n = 120$ rats and is provided freely in the `R`-based package `flare`. The response is represented by the expression level of the TRIM32 gene for all rats. This was the only gene that was found in the original study, to have a strong positive correlation with eight different genes that were known to cause a disorder known as Bardet-Biedl syndrome (BBS) which has vision loss as one of its major features. Upon inspection of the data, one rat displayed gene expression levels that were much different than the rest of the sample and has been excluded from the analysis.

The final genetic data and the correlation structure between the genes is presented in Figure 5 and it indicates that some genes have expression levels consistently higher that others, but at the same time it illustrates that some genes are highly correlated and thus summarizing this information under the form of a conditional independence graph brings additional information.

We evaluate GraphSIR, GraphSAVE and LassoSIR with respect to in-sample mean square error (MSE) and leave-one-out (LOO) mean square prediction error. The evaluation strategy goes as follows. We fit first GraphSIR, GraphSAVE and LassoSIR to the data, retain the $\hat{\boldsymbol{\beta}}_d$ coefficients for each direction and standardize them to have unit norm. We then define for each case i, a set of D new features denoted as $z_{1,i} = \boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}_1, \ldots, z_{D,i} = \boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}_D$ which we use as predictors in a linear regression to model the conditional expectation of the gene expression level. From the fitted regression model we retain next the estimated vector of coefficients denoted as $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_D)^\mathsf{T}$ and define the two evaluation criteria as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{z}_i^\mathsf{T}\hat{\boldsymbol{\gamma}})^2,$$
$$\text{LOO-cv} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{z}_{(-i)}^\mathsf{T}\hat{\boldsymbol{\gamma}}_{(-i)}\right)^2,$$

where y_i and $\boldsymbol{z}_i = (z_{1,i}, \ldots, z_{D,i})^\mathsf{T}$ represent for each case the response and the vector of new features obtained after performing SDR. The vector $\boldsymbol{z}_{(-i)} = (\boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}_{1(-i)}, \ldots, \boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}_{D(-i)})^\mathsf{T}$ represents the vector of values for the new features for case i obtained when the i-th case is excluded

from the training sample when performing SDR. Similarly, the vector $\hat{\boldsymbol{\gamma}}_{(-i)}$ represents the vector of estimated coefficients when the i-th case is excluded when fitting the regression model.

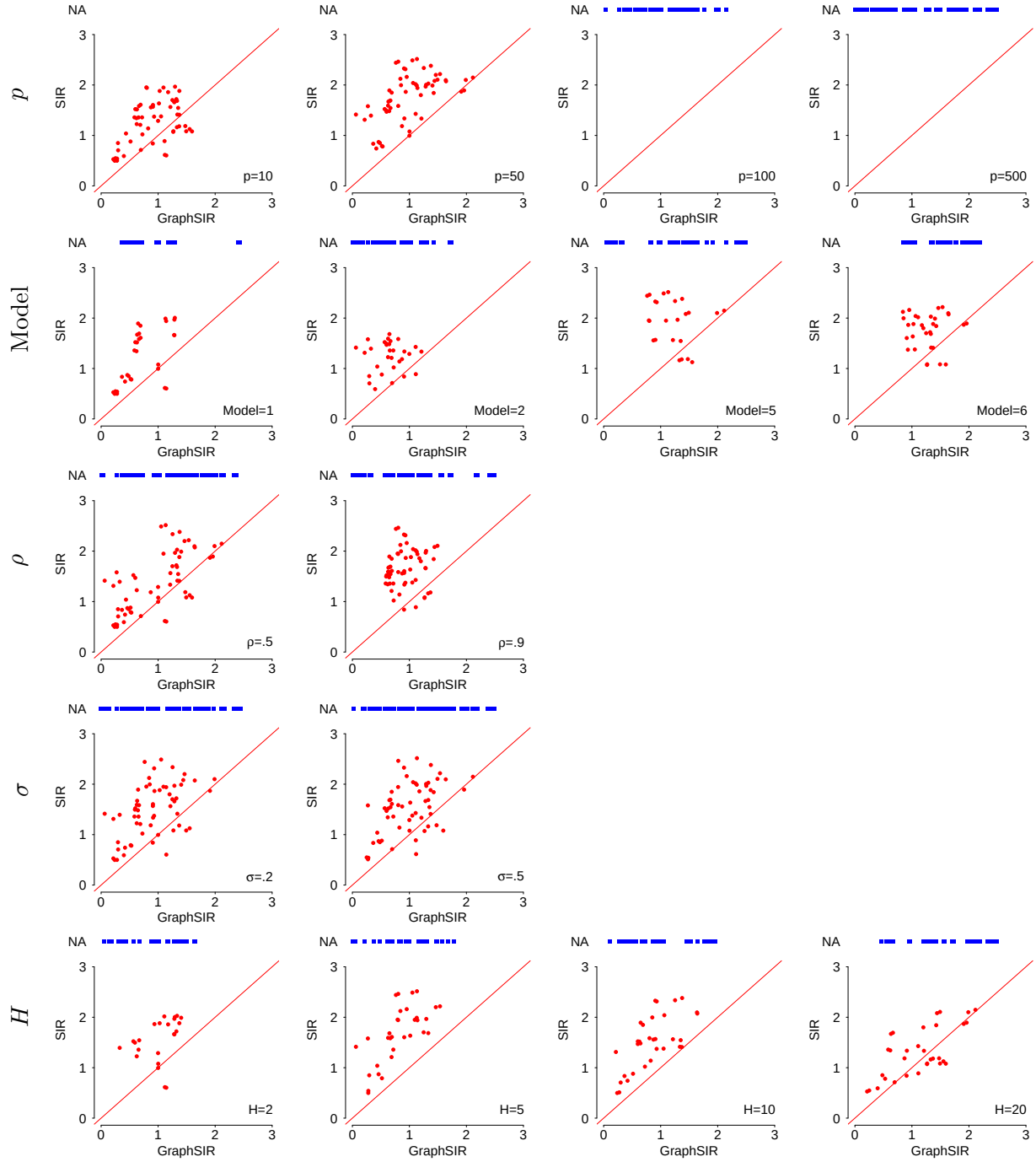

Figure 3: Simulation data. Frobenius loss (smaller is better) between $\hat{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_0$ for 256 scenarios split by values of p, model, values of ρ, σ and H.

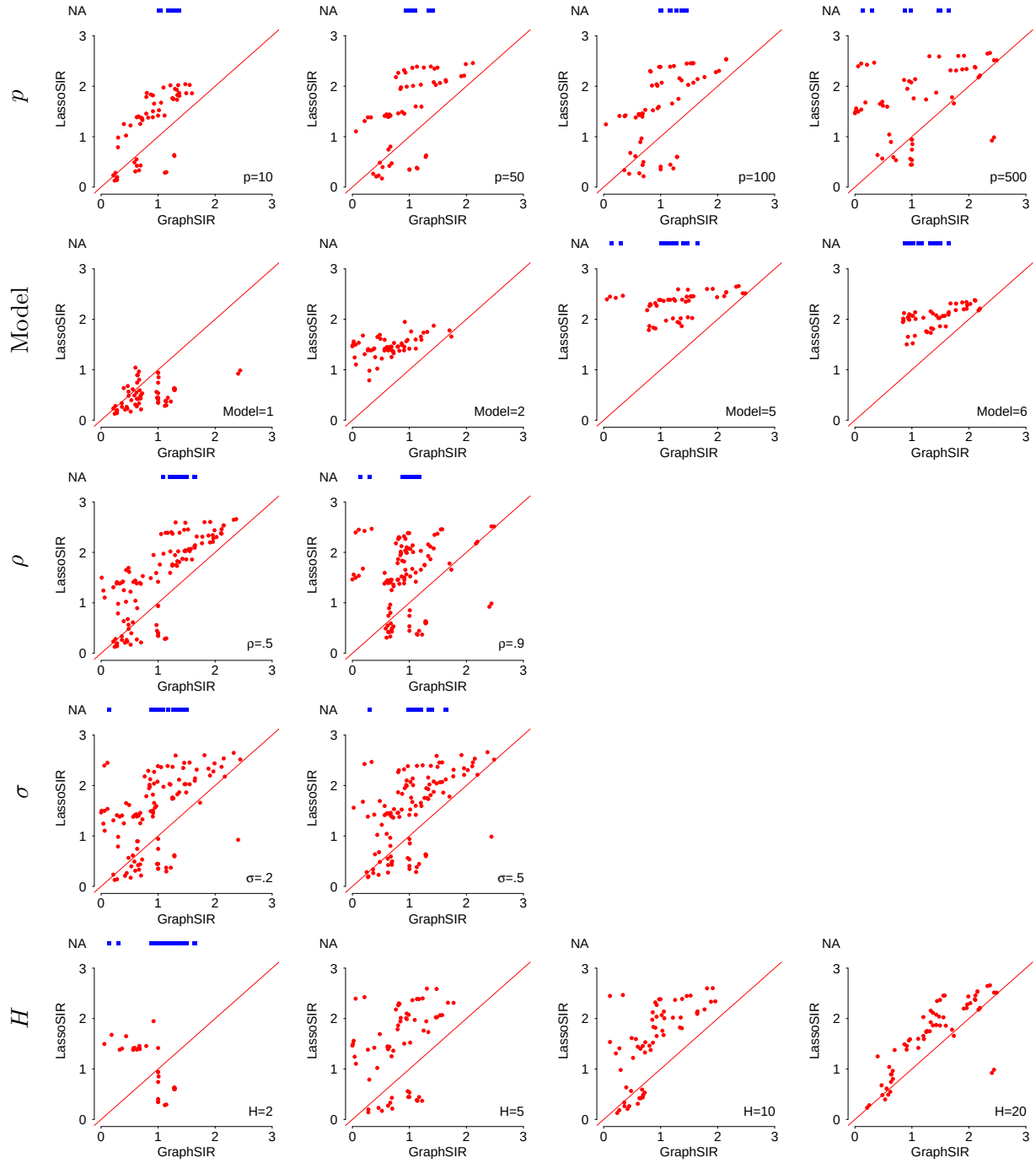

Figure 4: Simulation data. Frobenius loss (smaller is better) between $\hat{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_0$ for 256 scenarios split by values of p, model, values of ρ, σ and H.

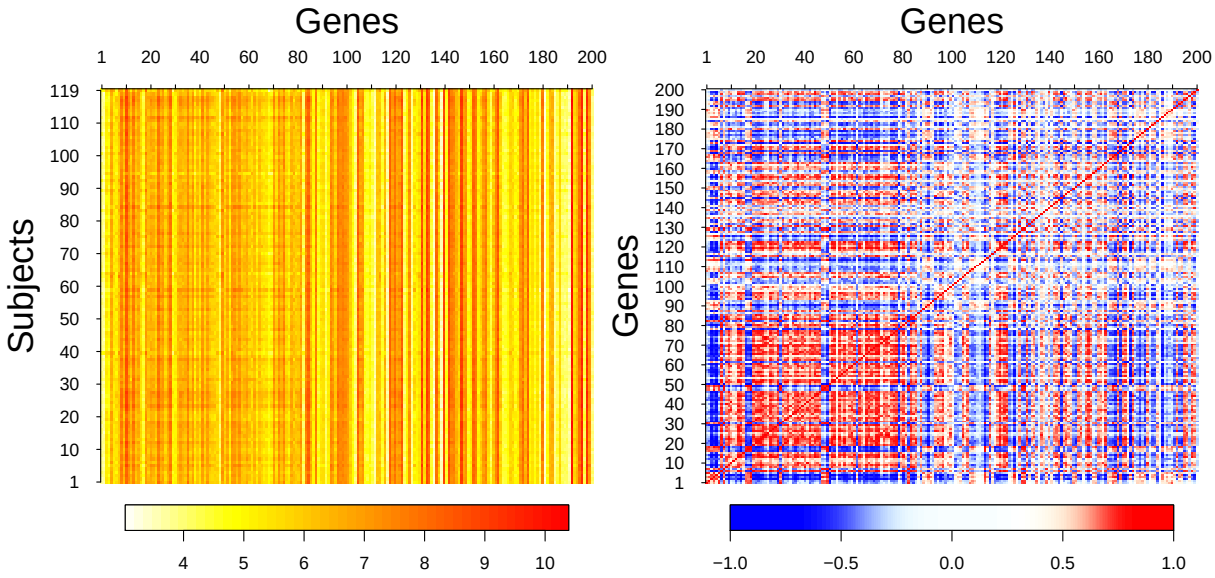

Figure 5: Eye data. Heatmap of the predictor gene expression levels for all subjects (left panel) and correlations between predictor genes (right panel).

Table 1 presents the obtained results when $D \in \{1, \ldots, 6\}$ and $H \in \{2, \ldots, 6\}$ and it illustrates that all three techniques provide low and very similar errors for the in-sample and out-of-sample predictions. Note that as is the case for classical SIR also LassoSIR suffers from the inability to estimate more than directions than slices, but no such problem is encountered for GraphSIR and GraphSAVE.

Table 1: Eye data: MSE$\times 10^2$ (top panels) and LOO-cv$\times 10^2$ (bottom panels) error for GraphSIR, GraphSAVE and LassoSIR for $2-6$ slices H with $1-6$ directions D.

	GraphSIR					GraphSAVE					LassoSIR				
D \ H	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
1	.83	.78	.73	.78	.78	.73	.73	.73	.73	.73	.59	.57	.54	.53	.53
2	.80	.73	.73	.73	.73	.73	.73	.73	.72	.73	.56	.56	.54	.54	.52
3	.56	.72	.73	.71	.72	.71	.72	.72	.72	.72	/	.54	.53	.52	.50
4	.57	.56	.72	.68	.71	.70	.71	.72	.71	.72	/	/	.53	.52	.49
5	.54	.54	.54	.68	.70	.68	.70	.72	.71	.68	/	/	/	.51	.49
6	.53	.53	.53	.54	.70	.62	.67	.72	.70	.65	/	/	/	/	.47
1	.91	.81	.87	.97	.90	.76	.75	.87	.75	.76	.79	.73	.81	.77	.76
2	.95	.91	.83	.92	.86	.76	.78	.77	.77	.78	.86	.73	.78	.73	.76
3	.92	.87	.88	.88	.88	.79	.78	.78	.78	.77	/	.76	.83	.74	.77
4	.94	.85	.90	.90	.90	.79	.78	.79	.79	.77	/	/	.86	.75	.79
5	.96	.85	.92	.88	.94	.78	.80	.79	.78	.74	/	/	/	.79	.76
6	.93	.85	.95	.89	.94	.76	.85	.79	.78	.75	/	/	/	/	.74

In Figure 6 the estimated sparse graph is plotted in panel (a) alongside scatterplots of the observed versus fitted values in panels (b)-(f) using GraphSIR with three slices and five

directions. This choice was made to illustrate that the proposed method produces meaningful results even when $D > H$. Panel (a) suggests that the conditional independence gene graph is relatively sparse, but identifies a group of interconnected genes that are mostly linked to a central, 'hub' gene. Panels (b)-(f) indicate that all five estimated directions tend to be positively correlated with the observed responses, thus indicating a good fit to the data. As a global summary, we fitted next the regression model of the response, the expression level of gene TRIM32, on the five selected directions and the obtained R^2 measure was 46%. A further residual analysis showed no serious deviations were the usual assumptions.

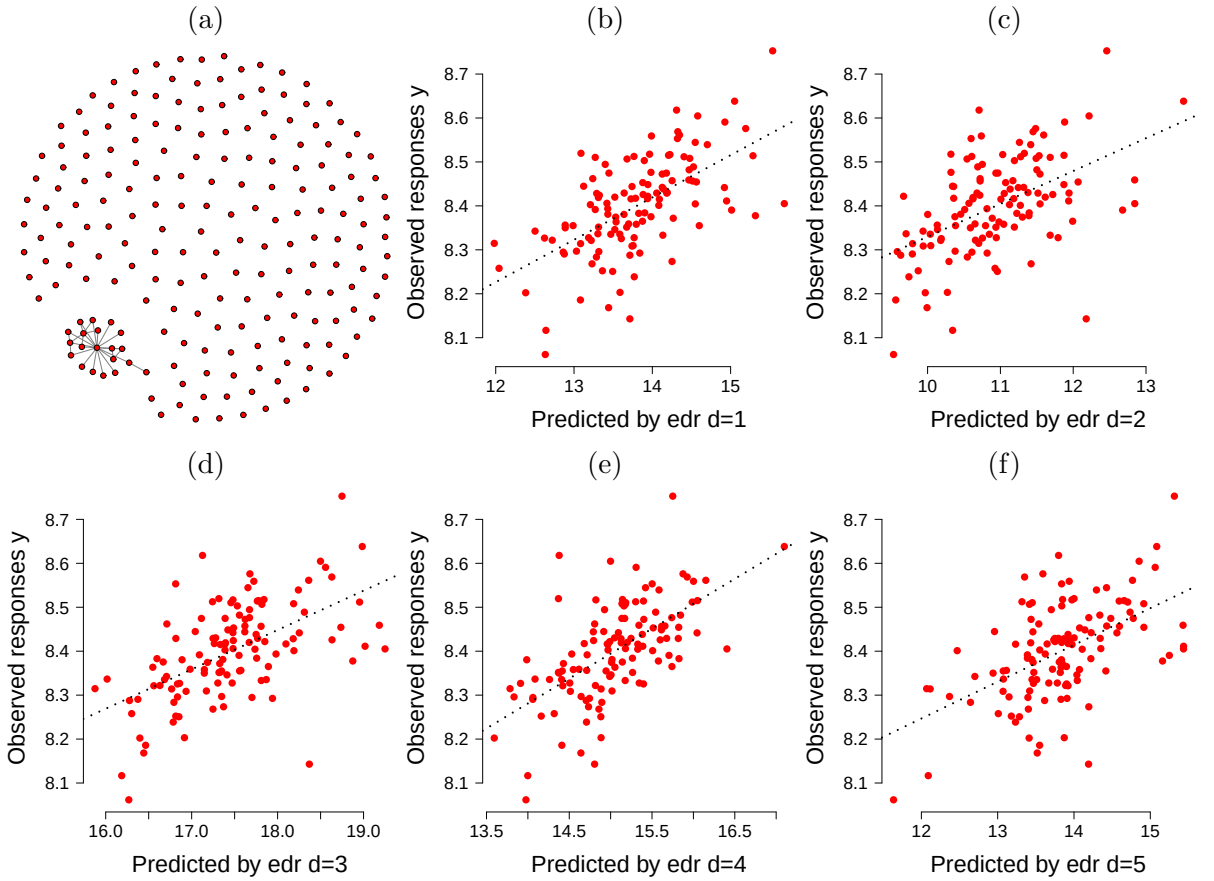

Figure 6: Eye data. The estimated graph obtained from $\hat{\mathbf{\Theta}}_X$ in panel (a) and scatterplots of the observed values y versus predicted values from GraphSIR with $H = 3$ and $D = 5$ in panels (b)-(f).

As we have mentioned in Section 1, the graph based procedures can be used also when the slicing variable is binary. To this end, we analyze the breast cancer data used in Augugliaro et al. (2013) where the purpose is to use genetic information to discriminate between subjects having cancer or not. The sample size is 52 and the number of genetic features is 287. The evaluation is very similar to the previous analysis: we fit GraphSIR and LassoSIR to the data, retain the $\hat{\boldsymbol{\beta}}_d$ coefficients and then define for each case i, a set of D new features denoted as $z_{1,i} = \boldsymbol{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_1, \ldots, z_{D,i} = \boldsymbol{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}_D$ which we use as predictors in a logistic regression. With the fitted model, we estimate next the probability of belonging to the cancerous group or not and evaluate the two competitors using ROC curves. The obtained results are presented in Figure 7 and we stress that LassoSIR cannot estimate more than one effective direction for this example, while GraphSIR does not suffer from this shortcoming. Moreover, the figure indicates that using one

direction in this case leads to suboptimal classification results and that the GraphSIR method is able to pick up more signal in the data due to the possibility of estimating more directions than available slices.

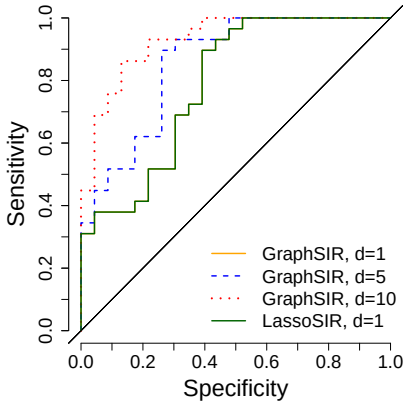

Figure 7: Breast cancer data. Estimated ROC curves for GraphSIR with $d = 1, 5$ or 10 and LassoSIR with $d = 1$ using a binary slicing variable. The solutions using $d = 1$ overlap in the figure.

7 Discussion

We present in this manuscript a high-dimensional version of slice inverse regression and an extension to the high-dimensional version of slice average variance estimation. Both approaches are rooted in a graph informed context, where high-dimensional graphs are used to extend the classical versions of SIR and SAVE.

The application on simulated data reveals that there is a substantial gain to be made by using the graph informed versions even for low-dimensional settings, where the performance was similar or better than that of the competitors, while the application to real data reveals once more that the graph based SIR is a worthy competitor to the high-dimensional LassoSIR.

Other extensions to other SDR techniques are as well possible in this framework, but not pursued here. The most obvious one would be to allow for a convex combination between the estimated matrices needed for GraphSIR and GraphSAVE as was proposed in Zhu et al. (2007). Instead of the ℓ_1 penalty, one could use a grouping or a fused penalty for the edge estimation as in Danaher et al. (2014) and Pircalabelu et al. (2016). If the multivariate normality of the vector $\boldsymbol{X}$ is considered too strong, then one can optimize the D-trace loss as in Zhang and Zou (2014), and if sparsity is desired at the level of the directions as well, then one can use a different constraint similar to what Molstad and Rothman (2018) propose. All of these extensions fall into the category of graph informed SDR and are directly applicable in the current framework.

References

Augugliaro, L., Mineo, A. M., and Wit, E. C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society: Series B*, 75(3):471–498.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, 102:130–142.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86(414):328–332.

Danaher, P., Wang, P., and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2):373–397.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 604–612. (NIPS).

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hilafu, H. and Yin, X. (2017). Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics*, 26(1):26–34.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278.

Li, B., Artemiou, A., and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414):316–327.

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(420):1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613.

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(2):580–610.

Lin, Q., Zhao, Z., and Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114(528):1726–1739.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Molstad, A. J. and Rothman, A. J. (2018). Shrinking characteristics of precision matrix estimators. *Biometrika*, 105(3):563–574.

Pircalabelu, E., Claeskens, G., and Waldorp, L. J. (2016). Mixed scale joint graphical lasso.

Biostatistics, 17(4):793–806.

Radchenko, P. (2015). High dimensional single index models. *Journal of Multivariate Analysis*, 139(C):266 – 282.

Ravikumar, P. D., Raskutti, G., Wainwright, M. J., and Yu, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of l_1-regularized MLE. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 1329–1336. (NIPS).

Scheetz, T., Kim, K.-Y., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V., and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Shin, S. J. and Artemiou, A. (2017). Penalized principal logistic regression for sparse sufficient dimension reduction. *Computational Statistics & Data Analysis*, 111:48 – 58.

Shin, S. J., Wu, Y., Zhang, H. H., and Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics & Data Analysis*, 52(9):4512 – 4520.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

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*, 64(3):363–410.

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 101(1):103–120.

Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474):630–643.

Zhu, L., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics & Data Analysis*, 51(5):2621 – 2635.

Zhu, L.-P. and Zhu, L. (2009). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis*, 100(5):862 – 875.