Biometrics 73, 529–539
June 2017

DOI: 10.1111/biom.12579

Structured Ordinary Least Squares: A Sufficient Dimension Reduction Approach for Regressions with Partitioned Predictors and Heterogeneous Units

Yang Liu,* Francesca Chiaromonte, and Bing Li*****

Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.
**email:* ywl5222@psu.edu
***email:* fxc11@psu.edu
***email:* bxl9@psu.edu

Summary. In many scientific and engineering fields, advanced experimental and computing technologies are producing data that are not just high dimensional, but also internally structured. For instance, statistical units may have heterogeneous origins from distinct studies or subpopulations, and features may be naturally partitioned based on experimental platforms generating them, or on information available about their roles in a given phenomenon. In a regression analysis, exploiting this known structure in the predictor dimension reduction stage that precedes modeling can be an effective way to integrate diverse data. To pursue this, we propose a novel Sufficient Dimension Reduction (SDR) approach that we call *structured Ordinary Least Squares* (sOLS). This combines ideas from existing SDR literature to merge reductions performed within groups of samples and/or predictors. In particular, it leads to a version of OLS for grouped predictors that requires far less computation than recently proposed groupwise SDR procedures, and provides an informal yet effective variable selection tool in these settings. We demonstrate the performance of sOLS by simulation and present a first application to genomic data. The R package "sSDR," publicly available on CRAN, includes all procedures necessary to implement the sOLS approach.

Key words: Data integration; Ordinary least squares; Structured data; Sufficient dimension reduction; Variable selection.

1. Introduction

High-dimensional data has become ubiquitous in many scientific fields, and *Sufficient Dimension Reduction* (SDR) is one way to overcome the challenges it poses for model selection and inference in regression problems (Li and Duan, 1989; Li, 1991; Cook and Weisberg, 1991; Adragni and Cook, 2009; Ma and Zhu, 2013). Considering the conditional distribution of a response $Y \in \mathbb{R}$ given a predictor vector $X \in \mathbb{R}^p$, SDR seeks a small number of linear combinations, that is, a low-dimensional linear subspace onto which to project X without loss of information on the regression. In symbols, it targets a dimension reduction subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$ ($\perp\!\!\!\perp$ indicates independence and $P_{\mathcal{S}}$ the orthogonal projection on $\mathcal{S}$). Naturally, the focus is on the smallest such $\mathcal{S}$, called the *central subspace* and denoted by $\mathcal{S}_{Y|X}$. Under mild conditions this is the intersection of all dimension reduction subspaces, which is minimal, unique, and represents the identifiable "parameter" of SDR (Cook, 2004).

While the central subspace captures all aspects of the conditional distribution of $Y|X$, some analyses focus on the mean function $\mathrm{E}(Y|X)$. Cook and Li (2002) proposed to consider mean dimension reduction subspaces $\mathcal{S}$ such that $\mathrm{E}(Y|X) = \mathrm{E}(Y|P_{\mathcal{S}}X)$ and their intersection, the minimal and unique *central mean subspace* $\mathcal{S}_{\mathrm{E}(Y|X)}$. It is easy to show that $\mathcal{S}_{\mathrm{E}(Y|X)} \subseteq \mathcal{S}_{Y|X}$ with equality for location regressions where $Y \perp\!\!\!\perp X | \mathrm{E}(Y|X)$.

Many methods exist for estimating $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{\mathrm{E}(Y|X)}$. For example, whatever the dimension of $\mathcal{S}_{\mathrm{E}(Y|X)}$, the direction spanned by the *Ordinary Least Squares* (OLS) vector falls within this space (and hence within $\mathcal{S}_{Y|X}$) (Li and Duan, 1989) under the so-called linearity condition. This requires $\mathrm{E}(X|P_{\mathcal{S}_{\mathrm{E}(Y|X)}}X)$ to be a linear function of X, and is guaranteed if X has an elliptical distribution. Other methods utilize the "inverse" regression of X on Y, for example, *Sliced Inverse Regression* (SIR; Li, 1991), *Sliced Average Variance Estimation* (SAVE; Cook and Weisberg, 1991), and *Directional Regression* (DR; Li and Wang, 2007), all of which estimate directions within $\mathcal{S}_{Y|X}$ under certain conditions. Yet other methods, such as *Minimum Average Variance Estimation* (MAVE; Xia et al., 2002), estimate $\mathcal{S}_{\mathrm{E}(Y|X)}$ using nonparametric regression tools; these require fewer conditions but are computationally more expensive.

In their traditional formulation, most SDR methods treat all predictors and statistical units the same. However, predictors and/or statistical units can present group structures relevant for analysis and interpretation. Predictors may be generated by different experimental platforms or belong to different phenomenological domains; for example, Guo et al. (2015) describe a regression where, due to the lack of instrumental climate records prior to the 19th century, past global surface temperatures are reconstructed as a function of a large number of climate proxies that naturally fall into different groups, such as tree composites, tree rings, ice cores, cave deposits, lake sediments, and historical records. At the same time, statistical units may originate from distinct studies, data collection efforts or subpopulations.

As a motivating application, we consider data from Kuruppumullage Don et al. (2013), who studied four types of mutations affecting DNA: small insertions, small deletions, nucleotide substitutions, and repeat number alterations at microsatellite loci. Divergence rates for these mutations were estimated in 1Mb (one million base pairs) non-overlapping windows along the human genome using primate alignments of neutrally evolving DNA. Hidden Markov Models run on the rates produced a segmentation of the genome into six distinct divergence states. The study also associated the states with 37 quantitative genomic features derived from publicly available genome-wide annotations for each of the 1Mb non-overlapping windows. These features were partitioned into eight groups of biochemical proxies—for example, for chromatin structure, transcription, etc. Our aim is to understand whether the prevalence of non-coding functional elements in any given window depends on the divergence states and the naturally partitioned genomic features. These elements are genomic sequences that do not encode proteins but have a function, for example, modulating the transcription of protein coding sequences. Focusing on one of the best annotated among them, we take as response variable a measurement of coverage by transcription start sites from (ENCODE Project Consortium and others, 2012). Thus, we have a regression where statistical units (windows) have a group structure determined by the divergence states, and quantitative predictors (genomic features) have a group structure determined by the biochemical processes and contexts they proxy. In this type of settings, accounting for and exploiting group information can be critical. Composite predictors accounting for partitions of X may be more informative, easier to interpret, and may simplify the analysis by "weeding out" whole groups of predictors with weak explanatory power. At the same time, composite predictors accounting for heterogeneity among units may reveal commonalities and differences in the relationship between Y and X across subpopulations. In some cases, accounting for groups may also further reduce dimension (Chiaromonte et al., 2002; Li et al., 2003, 2010; Hilafu and Yin, 2013; Guo et al., 2015).

As advances in experimental and computing technologies produce data that, in addition to high dimensional, are increasingly complex and structured, the need for considering group information in SDR is ever more pressing. Beyond the application introduced above, contemporary Genomics research requires integrating data across different high-throughput experimental platforms, multiple studies, diverse experimental conditions and various classifications of genes—or more generally of DNA regions (Louie et al., 2007; Wu et al., 2012; Kuruppumullage Don et al., 2013; Gomez-Cabrero et al., 2014). Our overarching goal is to develop methodology that employs SDR as a means to both reduce and integrate this type of data. We do this building upon a number of prior efforts; traditional SDR methods have been extended to account for the subpopulation structure induced by categorical predictors (Chiaromonte et al., 2002; Li et al., 2003) and, separately, to incorporate prior knowledge on predictor groups (Naik and Tsai, 2005; Li, 2009; Li et al., 2010; Guo et al., 2015). Notably, most methods that utilize partitions of the predictors do *not* require independence or uncorrelation across predictor groups; their ability to tackle interdependent groups is what makes them particularly appealing in applications. Combining ideas from this literature, we create a comprehensive approach for what we call *structured data*, that is, data with naturally grouped predictors *and* units.

As we will see in Section 2, we focus on *structured OLS* (sOLS). While in some respects OLS is inferior to inverse regression or nonparametric SDR methods, it is the least computationally expensive and most informationally parsimonious because it only extracts one dominant direction—in our case, one for each combination of predictor and unit groups. If for some such combinations several directions are relevant for the response, sOLS will not be able to capture them all. However, since the applications we have in mind are high dimensional and often characterized by relatively small sample sizes, it may not be reasonable to attempt estimation of multiple directions per combination. Section 5 discusses possible extensions and the handling of under-sampled regressions where the sample size n is smaller than the number of predictors p. Section 3 describes the performance of sOLS in simulations. Section 4 illustrates our application of sOLS to the data from Kuruppumullage Don et al. (2013).

2. Structured Ordinary Least Squares

2.1. *Sufficient Dimension Reduction for Structured Data*

Here, we introduce the theoretical formulation of *Structured Sufficient Dimension Reduction* at the population level. Because of our focus on OLS methodology, we do this from the perspective of the mean function, with the associated notions of structured mean dimension reduction subspace and structured central mean subspace (see below). These spaces are defined to fully preserve location information while conforming to both predictor and unit groups. We combine the frameworks of *groupwise SDR* (e.g., Li et al., 2010; Guo et al., 2015), where the reduction is partitioned through an orthogonal decomposition of the predictor space, and *partial SDR* (e.g., Chiaromonte et al., 2002; Li et al., 2003), where the reduction is "informed" by the subpopulation structure induced by one or more categorical predictors. Note that in groupwise SDR, the partition of the predictor space is orthogonal in terms of the Euclidean inner product $a^T b$, *not* in terms of the predictors' covariance—predictors can be dependent as well as correlated across groups. Note also that in partial SDR the categorical predictors are *not* reduced and comprised in linear combinations along with the quantitative predictors, but shape their reduction.

Consider the regression $Y|(X, W)$, where $Y \in \mathbb{R}$ is the response, $X \in \mathbb{R}^p$ is the vector of quantitative predictors to be reduced, and $W \in \{1, \ldots, c\}$ labels the subpopulations—in the case of several categorical predictors, the c levels of W represent all combinations of their levels. Next, denoting with $\oplus$ the direct sum of subspaces, consider an orthogonal decomposition $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$ of the predictor space; $\mathbb{R}^p = \mathcal{S}_1 \oplus \cdots \oplus \mathcal{S}_g$. For example, if $p = 6$ predictors are grouped as (X_1, X_2, X_3, X_6) and (X_4, X_5), we set $\mathcal{S}_1 = \text{span}(e_1, e_2, e_3, e_6)$ and $\mathcal{S}_2 = \text{span}(e_4, e_5)$, where $e_i \in \mathbb{R}^6$ has the ith element $= 1$ and all others $= 0$ (here, again, $\mathcal{S}_1$ and $\mathcal{S}_2$ are orthogonal with respect to the Euclidean inner product, but all kinds

of dependencies may exist across the corresponding groups of predictors).

DEFINITION 1. *If there exist subspaces $\mathcal{F}_i \subseteq \mathcal{S}_i$, $i = 1, \ldots, g$ such that*

$$\mathrm{E}(Y|X;W) = \mathrm{E}(Y|P_{\mathcal{F}_1}X, \ldots, P_{\mathcal{F}_g}X; W) \tag{1}$$

then $\mathcal{F} = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_g$ is a structured mean dimension reduction subspace *with respect to the orthogonal decomposition $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$ and the categorical variable W.*

This includes all information in X conforming to the orthogonal decomposition G, which is useful for predicting the conditional mean $\mathrm{E}(Y|X;W)$ when considered along with the categorical W. To focus on the smallest such subspace we use a Lemma from Li et al. (2010).

LEMMA 1. *Suppose that $\mathcal{F} = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_g$ and $\mathcal{F}' = \mathcal{F}'_1 \oplus \cdots \oplus \mathcal{F}'_g$ with $\mathcal{F}_i \subseteq \mathcal{S}_i$ and $\mathcal{F}'_i \subseteq \mathcal{S}_i$, $i = 1, \ldots, g$. Then $\mathcal{F} \cap \mathcal{F}' = (\mathcal{F}_1 \cap \mathcal{F}'_1) \oplus \cdots \oplus (\mathcal{F}_g \cap \mathcal{F}'_g)$.*

By Lemma 1 and under mild conditions (Cook, 1998; Yin et al., 2008), the intersection of all subspaces satisfying (1) still satisfies (1). A sufficient condition to guarantee this is as follows:

C.1 (support) for each $w = 1, \ldots, c$, the support of X_w is open and convex.

We are thus justified in making the following definition.

DEFINITION 2. *Under condition C.1, the* structured central mean subspace *with respect to the orthogonal decomposition $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$ and the categorical variable W is the intersection of all structured mean dimension reduction subspaces with respect to G and W. It is itself a structured mean dimension reduction subspace, denoted as $\mathcal{S}^{(G,W)}_{E(Y|X)}$.*

This setup mimics the one used to define the central mean subspace in traditional SDR literature. Similarly, one can mimic the setup used to define the central subspace, switching attention to the stronger condition $Y \perp\!\!\!\perp X|(P_{\mathcal{F}_1}X, \ldots, P_{\mathcal{F}_g}X; W)$. If this condition holds, $\mathcal{F} = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_g$ is a *structured dimension reduction subspace*; this includes all information in X which, conforming to G, is useful for predicting $Y|(X;W)$ along with W. Intersecting all such subspaces one obtains the *structured central subspace* $S^{(G,W)}_{Y|X}$.

In the remainder of this subsection, we present only one theorem which links $S^{(G,W)}_{\mathrm{E}(Y|X)}$ to groupwise central mean subspaces (Li et al., 2010) within subpopulations, and is critical for the development of the methodology proposed in the next subsection. The proof, along with an analogous theorem for $S^{(G,W)}_{Y|X}$ and more results connecting structured, groupwise and partial SDR are provided in the Supplement. Let $\mathcal{S}^{(G)}_{\mathrm{E}(Y_w|X_w)} = \oplus_{i=1}^{g}\mathcal{F}_{wi}$ be the groupwise central mean subspace within subpopulation w; its component spaces $\mathcal{F}_{wi} \subseteq \mathcal{S}_i$, $i = 1, \ldots, g$ are the smallest such that $\mathrm{E}(Y_w|X_w) = \mathrm{E}(Y_w|P_{\mathcal{F}_{w1}}X_w, \ldots, P_{\mathcal{F}_{wg}}X_w)$, where (X_w, Y_w) denotes a pair distributed as $(X, Y)|W = w$. We have:

THEOREM 1. *Under C.1, the structured central mean subspace with respect to G and W can be written as $\mathcal{S}^{(G,W)}_{\mathrm{E}(Y|X)} = \oplus_{w=1}^{c}\mathcal{S}^{(G)}_{\mathrm{E}(Y_w|X_w)} = \oplus_{w=1}^{c}\oplus_{i=1}^{g}\mathcal{F}_{wi} = \oplus_{i=1}^{g}\oplus_{w=1}^{c}\mathcal{F}_{wi}$.*

Of course $\mathcal{S}^{(G)}_{\mathrm{E}(Y_w|X_w)}$, $w = 1, \ldots, c$, can overlap in any fashion, but $\mathcal{S}^{(G,W)}_{\mathrm{E}(Y|X)}$ always coincides with their direct sum—which in turn can be reconstructed combining the subspaces $\mathcal{F}_{wi}$ first across the orthogonal decomposition and then over subpopulations, or conversely first over subpopulations for each predictor group and then across the orthogonal decomposition. In the equations above, we are using the symbol $\oplus$ both to indicate the direct sum of orthogonal subspaces when adding across predictor groups $i = 1, \ldots, g$, and that of subspaces that are not required to be orthogonal when adding over subpopulations $w = 1, \ldots, c$ (later in the article, we sometimes use the word "stacking," as opposed to the more general "combining," to distinguish in an intuitive way the direct sum of orthogonal subspaces). Importantly though, what Theorem 1 tells us is that $\mathcal{S}^{(G,W)}_{\mathrm{E}(Y|X)}$ can be estimated from subspaces obtained performing groupwise SDR within subpopulations.

As an example, consider again $X \in \mathbb{R}^6$ partitioned in $\mathcal{S}_1 = \mathrm{span}(\gamma_1)$ and $\mathcal{S}_2 = \mathrm{span}(\gamma_2)$, where $\gamma_1 = (e_1, e_2, e_3, e_6)$ and $\gamma_2 = (e_4, e_5)$. Along with X, consider a categorical $W \in \{1, 2\}$ that labels units from two subpopulations. Suppose the true regression models in the two subpopulations are: $Y_1 = \sin(X_{11} + X_{12} - X_{13} - X_{16}) + \cos(X_{14} + X_{15}) + \varepsilon_1$, $Y_2 = \sin(-X_{21} - X_{22} + X_{23} + X_{26}) + \cos(X_{24} - X_{25}) + \varepsilon_2$, with independent additive errors. Here, within each subpopulation, each group of predictors affects the response through a single direction. For $w = 1$ we have $\beta_{11} = (1, 1, -1, -1)^T$ and $\beta_{12} = (1, 1)^T$, while for $w = 2$ we have $\beta_{21} = (-1, -1, 1, 1)^T$ and $\beta_{22} = (1, -1)^T$. Consequently, the groupwise central mean subspaces within the two subpopulations are $\mathcal{S}^{(G)}_{\mathrm{E}(Y_1|X_1)} = \mathcal{F}_{11} \oplus \mathcal{F}_{12} = \mathrm{span}(\gamma_1\beta_{11}) \oplus \mathrm{span}(\gamma_2\beta_{12})$ and $\mathcal{S}^{(G)}_{\mathrm{E}(Y_2|X_2)} = \mathcal{F}_{21} \oplus \mathcal{F}_{22} = \mathrm{span}(\gamma_1\beta_{21}) \oplus \mathrm{span}(\gamma_2\beta_{22})$, and the structured central mean subspace is $\mathcal{S}^{(G,W)}_{\mathrm{E}(Y|X)} = \mathcal{S}^{(G)}_{\mathrm{E}(Y_1|X_1)} \oplus \mathcal{S}^{(G)}_{\mathrm{E}(Y_2|X_2)}$ which, combining the $\mathcal{F}_{wi}$'s, is simply the 3-dimensional span of the vectors $(1, 1, -1, 0, 0, -1)^T$, $(0, 0, 0, 1, 0, 0)^T$ and $(0, 0, 0, 0, 1, 0)^T$.

In actual applications the γ_i's will be known and fixed, while the β_{wi}'s will need to be estimated. Note also that, although in this example the dimension contributed by each group of predictors within each subpopulation is 1, in full generality it could be larger—possibly with spans that overlap across subpopulations. However, as discussed in the Introduction, our OLS-based methodology does in fact rely on the notion that each predictor group within each subpopulation is adequately summarized by at most one direction.

2.2. *Proposed Methodology*

Based on the above theoretical formulation, we now describe novel methodology for structured SDR. From Theorem 1, this is naturally organized into an *inner level*, which accomplishes groupwise SDR within each subpopulation, and an *outer level*, which combines the resulting subspaces across subpopulations—akin to partial SDR. While

the *inner level* could utilize any existing groupwise SDR method (e.g., *assembled Sliced Inverse Regression* (aSIR; Li, 2009), *groupwise Minimum Average Variance Estimation* (gMAVE; Li et al., 2010), *groupwise Sliced Inverse Regression* (gSIR; Guo et al., 2015), *groupwise Directional Regression* (gDR; Guo et al., 2015)), we focus on Ordinary Least Squares for the reasons explained earlier.

Our *groupwise OLS* (gOLS) targets directions relevant for the mean function $\mathrm{E}(Y|X)$ while accounting for a partition of the predictors expressed by the orthogonal decomposition $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$ of $\mathbb{R}^p$. In other words, it targets directions in the groupwise central mean subspace $\mathcal{S}^{(G)}_{\mathrm{E}(Y|X)}$ (Li et al., 2010). Following the *direct sum envelope approach* of Guo et al. (2015), the main idea is to enclose the subspace targeted by a traditional SDR method—in this case $\mathcal{S}_{\mathrm{E}(Y|X)}$ estimated by OLS—with a subspace that conforms to the group structure.

Since by definition any groupwise dimension reduction subspace is also a traditional dimension reduction subspace, one has $\mathcal{S}_{\mathrm{E}(Y|X)} \subseteq \mathcal{S}^{(G)}_{\mathrm{E}(Y|X)}$. That is, the traditional central mean subspace, which preserves information on $\mathrm{E}(Y|X)$, is bound to be contained in the groupwise central mean subspace, which preserves information on $\mathrm{E}(Y|X)$ and at the same time conforms to the given predictor groups. Another way of thinking of this is that accounting for predictor groups limits the reduction one can perform on X.

The direct sum envelope approach of Guo et al. (2015) allows one to find the smallest subspace that both conforms to the orthogonal decomposition G and encloses the subspace spanned by the columns of a $p \times r$ random matrix U. In particular, we use the following definition:

Definition 3. *Given an orthogonal decomposition* $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$ *of* $\mathbb{R}^p$ *and a subspace* $\mathcal{S}$ *of* $\mathbb{R}^p$*, consider the collection of subspaces*

$$\mathfrak{A} = \{\mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_g : \mathcal{H}_1 \subseteq \mathcal{S}_1, \ldots, \mathcal{H}_g \subseteq \mathcal{S}_g, \\ \mathcal{S} \subseteq \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_g \text{ almost surely}\}.$$

The intersection of all members of $\mathfrak{A}$ *is called the direct-sum envelope of* $\mathcal{S}$ *with respect to* G*, and is denoted by* $\mathcal{E}^{\oplus}(\mathcal{S}|G)$.

Thus, $\mathcal{E}^{\oplus}(\mathcal{S}|G)$ is the unique and smallest subspace that both conforms to G and encloses $\mathcal{S}$. Under C.1, if $\mathcal{S} \subseteq \mathcal{S}_{\mathrm{E}(Y|X)}$, then $\mathcal{E}^{\oplus}(\mathcal{S}|G) \subseteq \mathcal{S}^{(G)}_{\mathrm{E}(Y|X)}$, where equality holds if and only if $\mathcal{S}$ is not a proper subspace of $\mathcal{S}^{(G)}_{\mathrm{E}(Y|X)}$. This property of the direct sum envelope suggests how to derive an estimator for $\mathcal{S}^{(G)}_{\mathrm{E}(Y|X)}$ from any estimator for $\mathcal{S}_{\mathrm{E}(Y|X)}$ and provides the condition required to guarantee *exhaustiveness* (i.e., recovering of the whole space of interest) as defined by Li et al. (2005).

In the context of gOLS, $\mathcal{S}$ spanned by a single vector in $\mathbb{R}^p$, representing the overall OLS direction. This has a particularly attractive consequence; namely, that gOLS can be solved explicitly without resorting to the iterative least-squares optimization required by other groupwise SDR estimates, such as those proposed in Li et al. (2010) and Guo et al. (2015). This is due to the following theorem (a proof is provided in the Supplement).

Theorem 2. *Let* $G = (\mathcal{S}_1, \ldots, \mathcal{S}_g)$ *be an orthogonal decomposition of* $\mathbb{R}^p$*,* $v \in \mathbb{R}^p$ *and* $\mathcal{S} = \mathrm{span}(v)$*. Then*

$$\mathcal{E}^{\oplus}(\mathcal{S}|G) = \mathrm{span}(P_{\mathcal{S}_1} v) \oplus \cdots \oplus \mathrm{span}(P_{\mathcal{S}_g} v). \qquad (2)$$

As a result, the direct-sum envelope of the overall OLS vector can be obtained explicitly by projecting it onto the subspaces corresponding to each group.

Once gOLS has produced estimates of b_{wi}, $i = 1, \ldots, g$, for each subpopulation $w = 1, \ldots c$ at the inner level, we combine at the *outer level* the corresponding directions across subpopulations based on Theorem 1. First, for each predictor group we utilize an eigen decomposition to combine directions over the c subpopulations. Then, we "stack" the resulting spaces across the g predictor groups, which are orthogonal by construction. Importantly, the eigen decompositions order directions in each predictor group based on the magnitudes of their eigenvalues, which is relevant for dimension estimation (see below). In summary, our sOLS performs structured sufficient dimension reduction with a two-level procedure employing gOLS at the inner level, and eigen decompositions at outer level.

As mentioned in the Introduction, for OLS to produce a direction within the space of interest, the linearity condition must be satisfied. In our context this means:

C.2 (linearity) For each $w = 1, \ldots, c$, $E(X_w | P_{\mathcal{S}^{(G)}_{E(Y_w|X_w)}} X_w)$ is linear in X_w.

Versions of the linearity condition similar to C.2 are used by many SDR methods, especially those that are computationally simple. In many situations this does not impose a serious restriction, because the predictors are often pre-transformed to approximate normality, and also because low-dimensional projections of high-dimensional random vectors tend to have a normal distribution (Hall and Li, 1993).

2.3. *Numerical Implementation of sOLS*

Let (X_j, W_j, Y_j), $j = 1, \ldots, n$, be an i.i.d. sample of size n from the joint distribution of (X, W, Y). To explicitly distinguish the subpopulations labeled by $w = 1, \ldots, c$, we use the notation $(X_w, Y_w) \sim (X, Y)|W = w$ for predictor vector and response within subpopulation w, and $(X_{w,j}, Y_{w,j})$, $j = 1, \ldots, n_w$ for the corresponding subsample, with $n = \sum_{w=1}^{c} n_w$. Without loss of generality, we assume the predictor vector to be centered in each subpopulation, and estimate its $p \times p$ covariance matrix as $\hat{\Sigma}_w = n_w^{-1} \sum_{j=1}^{n_w} X_{w,j} X_{w,j}^T$.

Concerning the predictor groups, we use γ_i, $i = 1, \ldots, g$ to indicate known $p \times p_i$ matrices spanning the subspaces of the orthogonal partition $G = \{\mathcal{S}_1, \ldots, \mathcal{S}_g\}$. Without loss of generality, we assume $\gamma_i^T \gamma_i = I_{p_i}$ for all $i = 1, \ldots, g$. Accordingly, within each subpopulation $w = 1, \ldots, c$ and for each predictor group $i = 1, \ldots, g$ we use the notation $X_{wi} = \gamma_i^T X_w$ for the restricted predictor vector and $X_{wi,j}$, $j = 1, \ldots, n_w$ for its observations. The restricted $p_i \times p_i$ covariance matrix is estimated as $\hat{\Sigma}_{wi} = n_w^{-1} \sum_{j=1}^{n_w} X_{wi,j} X_{wi,j}^T$.

Finally, we let d_{wi}, $w = 1, \ldots, c$, $i = 1, \ldots, g$ indicate the dimension contributed within the wth subpopulation

by the ith predictor group, $d_w = \sum_{i=1}^{g} d_{wi}$, $d_i = \sum_{w=1}^{c} d_{wi}$ and $d = \sum_{i=1}^{g} d_i$. Recall our use of OLS implies that all d_{wi}'s are either 0 or 1; *we assume temporarily that they are all known.* Our numerical implementation proceeds as follows:

Inner level:

- Within each subpopulation $w = 1, \ldots, c$, estimate the overall $p \times 1$ OLS vector $\hat{b}_w = \hat{\Sigma}_w^{-1} n_w^{-1} \sum_{j=1}^{n_w} X_{w,j}^T Y_{w,j}$.
- Split $\hat{b}_w$ into $\hat{b}_{wi} = \gamma_i^T \hat{b}_w$, $i = 1, \ldots, g$.

Outer level:

- For each predictor group $i = 1, \ldots, g$, form $V_i = \sum_{w=1}^{c} (n_w/n) \hat{b}_{wi} \hat{b}_{wi}^T$ and compute its eigenvectors $v_{i1}, \ldots, v_{id_i}$ corresponding to the d_i largest eigenvalues, *assuming temporarily that the* $\hat{b}_{wi}$, $w = 1, \ldots, c$, *are linearly independent* (so that these eigenvalues are all > 0).
- Form the subspaces $A_i = \text{span}(v_{i1}, \ldots, v_{id_i}) \subseteq \mathcal{S}_i$.
- Estimate the structured central mean subspace of dimension d stacking such spans over the orthogonal partition of the predictors; that is, set $\hat{S}_{E(Y|X)}^{(G,W)} = \text{span}(A_1 \oplus \ldots \oplus A_g)$.

At the inner level, each $\hat{b}_{wi}$ is simply obtained from the ith predictor group components of the overall OLS vector estimated within the wth subpopulation. The outer level combines such $\hat{b}_{wi}$'s across subpopulations using an eigen decomposition—first separately for each predictor group, and then stacking over predictor groups. We also note that, even though the asymptotic distribution of the $\hat{b}_{wi}$'s is not directly utilized in our developments, it follows straightforwardly from that of the overall OLS $\hat{b}_w$: Let $b_w = E[(X_w - E(X_w))(Y_w - E(X_w))]$ and $\Lambda_w = E[(Y_w - E(Y_w))^2 (X_w - E(X_w))(X_w - E(X_w))^T]$. Then, as $n_w \to \infty$, we have $\sqrt{n_w}(\hat{b}_w - b_w) \xrightarrow{\mathcal{D}} N(0, \Lambda_w)$. Consequently, for each $i = 1, \ldots, g$, we have

$$\sqrt{n_w}(\hat{b}_{wi} - b_{wi}) \xrightarrow{\mathcal{D}} N\left(0, \gamma_i^T \Lambda_w \gamma_i\right).$$

Importantly, in the above description we have assumed that we know whether each d_{wi} is 0 or 1, and that for each predictor group i the vectors $\hat{b}_{wi}$, $w = 1, \ldots, c$, are linearly independent. In practice, we will have to determine whether or not each predictor group contributes a nontrivial direction within each subpopulation, and we will have to assess whether nontrivial directions for the same predictor group overlap across subpopulations—leading to smaller spans. This is the topic of the next subsection.

2.4. *Dimension Estimation*

Dimension estimation is needed at both the inner and outer levels. At the inner level, within each subpopulation w, we need to estimate the dimensions $(d_{w1}, \ldots, d_{wg})$ contributed by each predictor group. At the outer level, we need to estimate how these dimensions combine across subpopulations to produce the actual d_i for each predictor group: in the case where directions do not overlap, we simply have $d_i = \sum_{w=1}^{c} d_{wi}$ as assumed in the description of our numerical implementation—but in many practical applications we may very well have $d_i < \sum_{w=1}^{c} d_{wi}$. Finally, stacking over orthogonal predictor groups gives us the dimension of the structured central mean subspace, $d = \sum_{i=1}^{g} d_i$.

Inner level: In order to estimate the dimensions $(d_{w1}, \ldots, d_{wg})$ when applying gOLS within a subpopulation w, we propose a Bayesian Information Criterion (BIC) approach. Without loss of generality, assume that both the predictor and the response in the subsample $(X_{w,j}, Y_{w,j})$, $j = 1, \ldots, n_w$ are *standardized*—that is, they are centered to have mean 0 and rescaled by $\hat{\Sigma}_w^{-1/2}$ and $\sigma_{y,w}^{-1}$, respectively, where $\sigma_{y,w}$ is the standard deviation of $Y_{w,j}$. Let $(\hat{b}_{w1}, \ldots, \hat{b}_{wg})$ be the gOLS estimates produced by running our algorithm on the standardized data, using the working dimensions $\tilde{d}_{wi} = 1$, $i = 1, \ldots, g$. Without loss of generality, assume the $\hat{b}_{wi}$'s are ordered by decreasing norm: $||\hat{b}_{w1}|| \geq ||\hat{b}_{w2}|| \ldots \geq ||\hat{b}_{wg}||$. Our inner-level criterion is

$$G_{n_w}^{(IL)}(k) = \sum_{i=0}^{k} \lambda_i(M_w) - \frac{k+1}{n_w^{\phi} log(n_w)} \quad , \quad k = 0, 1, \ldots, g \qquad (3)$$

where $0 < \phi < 1/2$, $M_w = (\hat{b}_{w1} \oplus \ldots \oplus \hat{b}_{wg})(\hat{b}_{w1} \oplus \ldots \oplus \hat{b}_{wg})^T$, and $\lambda_i(M_w) = ||\hat{b}_{wi}||^2$ is its ith eigenvalue (recall the partition of the predictor space is orthogonal). Intuitively, as $n_w \to \infty$, the first term in $G_{n_w}^{(IL)}(k)$ increases as k increases, while the second is negligible before k reaches $d_w = \sum_{i=1}^{g} d_{wi}$, and dominant afterward. The net effect is that, for large n_w, the criterion tends to be maximized at the true dimension d_w. We thus estimate d_w by the integer $\hat{d}_w$ that maximizes the criterion (3), and set $\hat{d}_{wi} = 1$ for $i = 1, \ldots, \hat{d}_w$ and $\hat{d}_{wi} = 0$ for $i = \hat{d}_w + 1, \ldots, g$. The next theorem, which is proved in the Supplement, establishes that these estimated dimensions convergence in probability to the true dimensions. We assume that, for each subpopulation w and group i, the population-level OLS vector b_{wi} satisfies the following condition:

C.3 (coverage) $b_{wi} \neq 0$ whenever $\mathcal{F}_{wi} \neq \{0\}$.

In the SDR literature, this is known as the *coverage* condition: once the linearity condition guarantees that b_{wi} belongs to the space $\mathcal{F}_{wi}$, coverage is used to ensure that b_{wi} spans such space. In practice, we are eliminating from consideration the special situation in which, due to symmetry of the regression surface about the origin, estimators such as OLS cannot estimate the targeted central mean subspace (Cook and Li, 2002).

THEOREM 3. *If conditions C.1 ∼ C.3 are satisfied, then*

$$\lim_{n \to \infty} P[(\hat{d}_{w1}, \ldots, \hat{d}_{wg}) = (d_{w1}, \ldots, d_{wg})] = 1.$$

Note that if $\hat{d}_{wi} = 0$, the mean function $\text{E}(Y_w | X_w)$ within subpopulation w does not depend on any of the predictors belonging group i; this corresponds to $\hat{b}_{wi} = 0$ in the gOLS estimate.

Outer level: Having produced dimension estimates $(\hat{d}_{w1}, \ldots, \hat{d}_{wg})$ and gOLS estimates $(\hat{b}_{w1}, \ldots, \hat{b}_{wg})$ for each w, at the outer level we combine these vectors across subpopulations using an eigen decomposition for each predictor group, and then stacking over predictor groups. At the population level, some of the $b_{1i}, \ldots, b_{ci}$ for group i may be 0 vectors, and those that are not may exhibit linear dependence. The dimension d_i is simply the column rank of the matrix $(b_{1i}, \ldots b_{ci})$. We again use a BIC approach. As with the inner level BIC, we assume without loss of generality that predictor vector and response are standardized within subsamples. Our outer-level criterion is

$$G^{(OL)}_{n_{\min}}(k) = \sum_{\ell=1}^{k} \lambda_\ell(\hat{M}_i) - \frac{k}{n_{\min}^{\psi}} \quad , \quad k = 1, \ldots, c, \qquad (4)$$

where $0 < \psi < 1/2$, $n_{\min} = \min_w n_w$, $\hat{M}_i = (\hat{b}_{1i}, \ldots, \hat{b}_{Ci})(\hat{b}_{1i}, \ldots, \hat{b}_{Ci})^T$ and $\lambda_\ell(\hat{M}_i)$ is its ℓth eigenvalue. The next Theorem states that the estimate $\hat{d}_i$ obtained maximizing the criterion convergences in probability to the true dimension. The proof is given in the Supplement.

THEOREM 4. *If conditions C.1 ∼ C.3 are satisfied, then* $\lim_{n\to\infty} P(\hat{d}_i = d_i) = 1$.

In addition to the BIC approach, we can also estimate d_i by extending the bootstrap procedure of Ye and Weiss (2003). The underlying idea is that the average of the distances between the space estimated on the original sample and the spaces estimated on the bootstrap samples ought to be smallest in the vicinity of the true dimension d_i. The bootstrap procedure has the advantage of being entirely data-driven, not requiring any assumption, and especially not relying on large sample sizes. We resample the samples from each subpopulation separately, so as to maintain the subpopulation proportions, and proceed as follows:

- Based on $\hat{d}_{wi}$ and $\hat{b}_{wi}$, $w = 1, \ldots, c$, obtained by running inner-level dimension estimation and gOLS on the original sample, set $\tilde{d}_i = \sum_{w=1}^{c} \hat{d}_{wi}$, form the matrix $V_i = \sum_{w=1}^{c} \frac{n_w}{n} \hat{d}_{wi}(\hat{b}_{wi}\hat{b}_{wi}^T)$, and take the eigenvectors $v_{i1}, \ldots, v_{i\tilde{d}_i}$ of the $\tilde{d}_i$ largest eigenvalues of this matrix.
- Fix the dimension estimates $\hat{d}_{wi}$, $w = 1, \ldots, c$, from the original sample and, for the $m = 1, \ldots, M$ bootstrap samples, run the inner-level gOLS to obtain $\hat{b}^{(m)}_{wi}$, $w = 1, \ldots, c$. Form the matrix $V_i^{(m)} = \sum_{w=1}^{c} \frac{n_w}{n} \hat{d}_{wi}(\hat{b}^{(m)}_{wi}\hat{b}^{(m)T}_{wi})$ and take the eigenvectors $v_{i1}^{(m)}, \ldots, v_{i\tilde{d}_i}^{(m)}$ of the $\tilde{d}_i$ largest eigenvalues of this matrix.
- Compute $h(k) = \frac{1}{M}\sum_{m=1}^{M} \| P_k - P_k^{(m)} \|_{\mathcal{F}}, \quad k = 1, \ldots, \tilde{d}_i$, where $\| \cdot \|_{\mathcal{F}}$ is the Frobenius norm (the square root of the sum of all squared entries of the argument matrix), $P_k = \sum_{\ell=1}^{k} v_{i\ell} v_{i\ell}^T$ is the orthogonal projection onto $\text{span}(v_{i1}, \ldots, v_{ik})$, and $P_k^{(m)} = \sum_{\ell=1}^{k} v_{i\ell}^{(m)} v_{i\ell}^{(m)T}$ is the orthogonal projection onto $\text{span}(v_{i1}^{(m)}, \ldots, v_{ik}^{(m)})$, $m = 1, \ldots, M$. For each k, this measures the average distance between spans produced by original and bootstrap samples.
- Select $\hat{d}_i$ as the minimizer of $h(k)$.

Finally, since the last step of the outer level of our procedure consists simply of stacking spans over the orthogonal partition of the predictors, our estimate for the dimension of the whole structured central mean subspace $S^{(G,W)}_{\mathrm{E}(Y|X)}$ is simply $\hat{d} = \sum_{i=1}^{g} \hat{d}_i$. All procedures for numerical implementation and dimension estimation in sOLS are provided as part of the R package "sSDR," publicly available on CRAN.

3. Simulation Study

We conducted an extensive simulation study to investigate the empirical performance of our methodology under a variety of scenarios. In each scenario, we measured subspace estimation accuracy, computational efficiency, and accuracy of our inner-level and outer-level dimension estimation. Due to space constraints, we only highlight key findings here, relegating details on simulation scenarios, full results and more extensive comments to the Supplement.

Regarding the inner level of our procedure we find that, when predictor groups exist, both gOLS and *assembled OLS* (aOLS, which utilizes OLS vectors obtained separately for each predictor group instead of projections of the overall OLS vector onto the groups) have higher subspace estimation accuracy than the overall OLS computed ignoring groups. However, gOLS consistently and substantially outperforms aOLS because, unlike the latter, it guarantees unbiased estimation regardless of interdependencies among predictor groups (Tables S1 and S2 in the Supplement). In addition, although gOLS can identify at most one direction in each predictor group, we find that in scenarios where several relevant directions exist it does succeed in capturing the dominant one (Table S8 in the Supplement). Finally, compared to other groupwise SDR methods such as gSIR, gDR (Guo et al., 2015) and gMAVE (Li et al., 2010), we find that gOLS is more accurate and vasty less computationally expensive (as much as four orders of magnitude lower running times than gMAVE; Tables S14 and S15 in the Supplement).

Regarding the outer level of our procedure we find that, when both predictor groups and subpopulations exist, sOLS has higher subspace estimation accuracy and induces lower prediction error in models fitted after dimension reduction than other least-squares-based SDR techniques that ignore subpopulations (gOLS), or ignore predictor groups (partial OLS, or pOLS, along the lines in Li et al., 2003), or ignore both (overall OLS) (Table 1; Tables S3, S4, S5, and S6 in the Supplement). We also find that, when combining estimated directions across subpopulations, applying weights proportional to the sizes of the samples from each subpopulation (see Subsection 2.3) matters for subspace estimation accuracy and subsequent model prediction error; if the data is unbalanced, ignoring these weights hinders performance (Table S7 in the Supplement).

Regarding inner-level and outer-level dimension estimation, we find that our BIC approaches work reasonably well especially for large sample sizes, as to be expected given the asymptotic properties discussed in Subsection 2.4 (Tables 2 and 3; Tables S9 and S10 in the Supplement). The results presented here are based on setting both ϕ in (3) and ψ in (4) to 1/8, a value that provides satisfactory performance on finite simulated samples. Also, our outer-level bootstrap dimension

Table 1

Estimation accuracy and prediction error: means (standard deviations) of the distances between true and estimated structured central mean subspaces (upper table), and of the prediction errors calculated with 100 test samples (lower table), over 100 simulated data sets

Measurement	ρ	θ	NSR	Method	n = 50	n = 100	n = 500	n = 1000
Space distance	0.3	1	7%	BM	2.497 (0.098)	2.493 (0.099)	2.501 (0.098)	2.493 (0.087)
				OLS	1.808 (0.032)	1.770 (0.019)	1.740 (0.003)	1.736 (0.001)
				gOLS	1.619 (0.101)	1.523 (0.064)	1.439 (0.014)	1.426 (0.007)
				pOLS	1.482 (0.027)	1.445 (0.019)	1.419 (0.002)	1.416 (0.001)
				sOLS	0.655 (0.139)	0.442 (1.121)	0.192 (0.039)	0.136 (0.027)
Prediction error	0.3	1	7%	OLS	3.244 (0.242)	3.213 (0.850)	2.983 (0.193)	2.969 (0.239)
				gOLS	3.230 (0.248)	3.182 (0.737)	2.961 (0.187)	2.936 (0.224)
				pOLS	1.812 (0.233)	1.772 (0.958)	1.447 (0.196)	1.441 (0.232)
				sOLS	1.582 (0.199)	1.476 (0.825)	1.159 (0.111)	1.151 (0.129)

Simulation model with two predictor groups and two subpopulations: $Y_w = \exp(0.8\beta_{w1}^T X_{w1}) + 2\beta_{w2}^T X_{w2} + \theta\varepsilon_w$, $w = 1, 2$, where $\beta_{11} \neq \beta_{21}$, $\beta_{12} \neq \beta_{22}$, $X \sim N_p(0, R)$ and $\varepsilon_w \sim N(0, 1)$. ρ is the pairwise correlation coefficient in the compound covariance matrix postulated for the predictors, θ controls the error variance, and NSR represents the ratio between variance of the error (noise) and variance of the regression mean function (signal). n is the sample size used in simulations. BM, benchmark generated using random directions; OLS, no grouping; gOLS, grouped predictors; pOLS, grouped units; sOLS, grouped predictors and units. Distances between subspaces are measured by the Frobenius norm of the difference between their projection matrices (smaller distances correspond to better estimation). Prediction errors are computed fitting parametric models (regression for OLS and gOLS, ANCOVA for pOLS and sOLS) after dimension reduction. See Supplement for more details and results.

estimator has satisfactory performance (Table 3; Table S11 in the Supplement). The outer-level BIC and bootstrap procedures tackle dimension estimation very differently—utilizing eigenvalues and eigenvectors in the eigen decompositions, respectively. Our simulation results suggest that both can be effective. Finally, we find that weighting actually hinders dimension estimation at the outer level because it leads to downplaying dimensional contributions from poorly sampled subpopulations (Tables S12 and S13 in the Supplement).

4. Application to Genomic Data

As a first application, we analyze the data from Kuruppumullage Don et al. (2013) (see Introduction). For $n =$ 2556 1Mb non-overlapping windows along the human genome, we consider a vector of quantitative genomic features (X, with $p = 37$), a divergence state label (W, with $w = 1, 2, \ldots, 6$) and coverage by transcription start sites (Y). Windows are partitioned into six divergence states: IDS− ($w = 1$, depressed insertions, deletions and substitutions); IDS− − ($w = 2$, strongly depressed insertions, deletions and substitutions, located exclusively on chromosome X); I+ ($w = 3$, moderately enhanced insertions); IDS+ ($w = 4$, strongly enhanced insertions, deletions and substitutions, located preferentially near the telomeres of autosomes); M++ ($w = 5$, strongly enhanced microsatellite alterations); and DS+ ($w = 6$, moderately enhanced deletions and substitutions). See Table S16 in the Supplement. Genomic features are partitioned into eight biochemical proxy groups: replication ($i = 1$); transposition ($i = 2$); recombination ($i = 3$); chromatin structure ($i = 4$); transcription ($i = 5$); methylation ($i = 6$); slippage ($i = 7$); and repair ($i = 8$). See Table S17 in the Supplement. Our sOLS can be applied to reduce dimension while accounting for both group structures simultaneously, and an appropriate parametric model to express Y as a function of the partitioned X and of W can be developed and fit after X has been reduced.

A complicating issue is that, since the windows are consecutive along the genome, adjacent observations present autocorrelation. To mitigate it, we form 100 independent random subsamples of $n = 1000$ windows (out of 2556) and repeat our analysis on each. An important side effect of this strategy, however, is that for some divergence states the sample size is now rather modest compared to the number of quantitative predictors (see again Table S16 in the Supplement)—supporting the use of a parsimonious approach such as sOLS, which seeks no more than one relevant direction for each predictor group in each divergence state.

On a technical note, we apply square-root transformation to the response to regularize its distribution, and then standardize both the response and the quantitative predictors.

Table 2

BIC *for inner level dimension estimation with gOLS: Proportion of cases with correct dimensions selected, out of 100 simulated data sets*

Dimensions	ρ	θ	n = 50	n = 100	n = 500	n = 1000
(1, 1)	0.3	1	0.70	0.90	1.00	1.00
(1, 0)	0.3	1	0.88	0.96	1.00	1.00
(0, 1)	0.3	1	1.00	1.00	1.00	1.00
(0, 0)	0.3	1	0.54	0.86	1.00	1.00

Simulation model with two predictor groups restricted to subpopulation $w = 1$: $Y_1 = \exp(0.8\beta_{11}^T X_{11}) + 2\beta_{12}^T X_{12} + \theta\varepsilon_1$, where $X \sim N_p(0, R)$ and $\varepsilon_w \sim N(0, 1)$. ρ is the pairwise correlation coefficient in the compound covariance matrix postulated for the predictors and θ controls the error variance. n is the sample size used in simulations. $(1, 1)$: $\beta_{11} \neq 0$ and $\beta_{12} \neq 0$; $(1, 0)$: $\beta_{11} \neq 0$ and $\beta_{12} = 0$; $(0, 1)$: $\beta_{11} = 0$ and $\beta_{12} \neq 0$; $(0, 0)$: $\beta_{11} = 0$ and $\beta_{12} = 0$. Accurate inner level dimension estimation with the BIC is, relatively speaking, harder in the presence of directions that contribute non-linearly to the mean function, and hardest in the $(0, 0)$ case, that is, in the absence of signal. See Supplement for more details and results.

Table 3

BIC *and Bootstrap for outer level dimension estimation: Proportion of cases with correct dimensions selected, out of 100 simulated data sets (gOLS with fixed known dimensions at the inner level)*

Method	Predictor group	Dimension	ρ	θ	n = 50	n = 100	n = 500	n = 1000
BIC	First	1	0.3	1	1.00	1.00	1.00	1.00
	Second	2	0.3	1	1.00	1.00	1.00	1.00
Bootstrap	First	1	0.3	1	1.00	1.00	1.00	1.00
	Second	2	0.3	1	0.79	0.86	0.97	0.98

Simulation model with two predictor groups and two subpopulations: $Y_w = \exp(0.8\beta_{w1}^T X_{w1}) + 2\beta_{w2}^T X_{w2} + \theta\varepsilon_w$, $w = 1, 2$, where $\beta_{11} = \beta_{21}$ and $\beta_{12} \neq \beta_{22}$; thus here $(d_1, d_2) = (1, 2)$. As in Tables 1 and 2, $X \sim N_p(0, R)$ and $\varepsilon_w \sim N(0, 1)$. ρ is the pairwise correlation coefficient in the compound covariance matrix postulated for the predictors and θ controls the error variance. Accurate outer level dimension estimation with the Bootstrap is, relatively speaking, harder when different subpopulations contribute distinct directions. See Supplement for more details and results.

Moreover, while we do use sample size weighting for estimation of the structured central mean subspace, we perform outer level dimension estimation *without* weighting; since the number of windows belonging to various divergence states are very different, this prevents deterioration in dimension estimation (see Section 3 and Supplement).

We use the BIC in (3) with $\phi = 1/8$ to perform inner-level dimension estimation. Counts of "votes" for $d_{wi} = 1$ (versus 0) obtained for every $i = 1, \ldots, 8$ and $w = 1, \ldots, 6$ across the 100 subsamples, along with summary inferences based on majority votes, are reported in Table S18 in the Supplement. Interestingly, the only two strongly supported predictor groups are *chromatin structure* ($i = 4$) with 100% votes in all states, and *slippage* ($i = 7$) with very strong effects in the "cold" state IDS− (100% votes) and the "warm" states I+ (100% votes) and DS+ (86% votes), but only weak effects in the "cold" state IDS− − (27% votes) and the "hot" states IDS++ (20% votes) and M++ (32% votes). Based on these results, we completely remove from further analysis replication, transposition, recombination, transcription, methylation, and repair. Next, we use gOLS to estimate directions within diverge states, including only the chromatin structure group in all states and the slippage group in IDS−, I+, and DS+.

Passing to the outer level, we combine directions across divergence states using eigen decompositions. To perform outer-level dimension estimation, we use the BIC in (4) with $\psi = 1/8$. Results shown in Table S19 of the Supplement support the choice $d_i = 1$ for both $i = 4$ and $i = 7$; for the chromatin structure predictor group a unique direction appears to work as a sufficient summary for all six divergence states, and for the slippage predictor group a unique direction appears to work as a sufficient summary for IDS−, I+, and DS+. Our bootstrap procedure confirms these inferences (Figure S1 in the Supplement). Thus, we reduce our analysis to a 2-dimensional structured central mean subspace, with one direction proxying chromatin structure and the other proxying slippage. The former is expected to contribute to the explanation of coverage by transcription start sites in all divergence states, and the latter in IDS−, I+, and DS+.

Figure 1 shows average loadings for these two directions across the 100 random subsamples. Remarkably, a strong dominant feature emerges in each; the number of dnase1 hypersensitive sites for chromatin structure, and the number of microsatellite loci for slippage.

After performing dimension reduction with sOLS, we use standard model building and diagnostics tools to construct a parametric ANCOVA model for the response as a function of the two composite predictors and the divergence states. Table 4 reports performance for various models comprising linear and quadratic terms; the one we chose has an excellent performance, with an R^2 of 77.3%. Its form is

$$Y_1 = \mu_1 + \eta_{chr} X_{1,chr} + \eta_{slp} X_{1,slp} + \varepsilon_1$$
$$Y_3 = \mu_1 + \Delta\mu_3 + \eta_{chr} X_{3,chr} + \eta_{slp} X_{3,slp} + \varepsilon_3$$
$$Y_6 = \mu_1 + \Delta\mu_6 + \eta_{chr} X_{6,chr} + \eta_{slp} X_{6,slp} + \tau_6 X_{6,chr} X_{6,slp} + \varepsilon_6$$
$$Y_w = \mu_1 + \Delta\mu_w + \eta_{chr} X_{w,chr} + \varepsilon_w, \quad w = 2, 4, 5$$

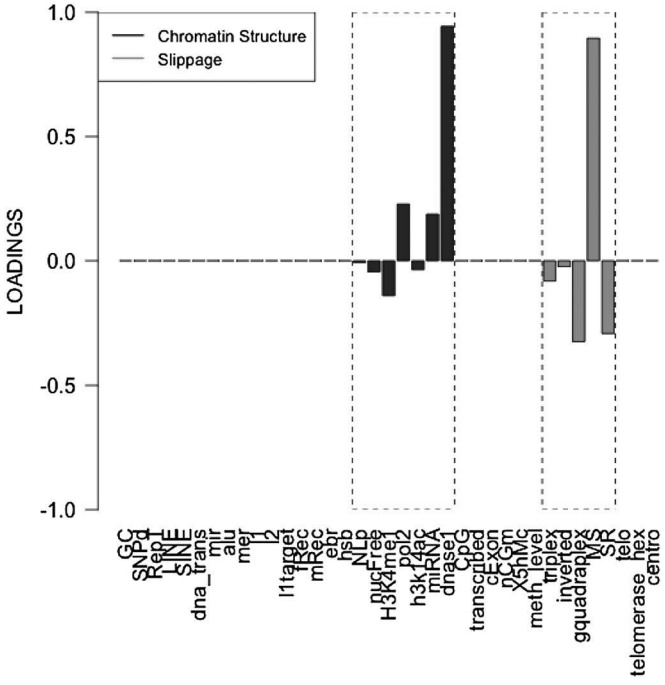

Figure 1. Relevant directions associated with chromatin structure and slippage predictor groups. Horizontal axis: 37 genomic features arranged by group. Vertical axis: average estimated loadings from 100 independent random subsamples of size $n = 1000$.

Table 4

Performance of ANCOVA models: means (standard deviations) of R^2 and prediction error over 100 *independent random subsamples of size $n = 1000$, and number of free parameters. Prediction error is computed associating to each subsample the test set of $\tilde{n} = 1556$ windows* not *included in it.*

ANCOVA model	R^2	Prediction error	Free parameters
Quadratic (all terms)	77.8% (1.0%)	0.053 (0.001)	27
Linear (all terms)	77.3% (1.0%)	0.053 (0.001)	15
Linear (only chromatin structure)	70.8% (1.2%)	0.059 (0.001)	12
Linear (only slippage)	8.6% (1.3%)	0.104 (0.001)	4
Final Model	77.3% (0.9%)	0.052 (0.001)	9

Quadratic (all terms): linear and quadratic terms in X_{chr} in all states; linear and quadratic terms in X_{slp} and interaction $X_{chr} \cdot X_{slp}$ in states IDS−, I+, and DS+. *Linear (all terms)*: linear terms in X_{chr} in all states; linear terms in X_{slp} in states IDS−, I+, and DS+. *Linear (only chromatin structure)*: linear terms in X_{chr} in all states. *Linear (only slippage)*: linear terms in X_{slp} in states IDS−, I+, and DS+. All models include the base intercept and intercept difference terms.

where Y_w, $X_{w,chr}$, $w = 1, 2, 3, 4, 5, 6$ and $X_{w,slp}$, $w = 1, 3, 6$ indicate the response and the (common) chromatin structure and slippage composite predictors as observed in different divergence states. This model has intercepts that vary among divergence states, parameterized using differences versus the intercept in IDS− (state $w = 1$). However, it has a shared linear slope for the chromatin structure composite predictor across all states, and a shared linear slope for the slippage composite predictor across the three states where such predictor is relevant. Moreover, the model comprises only one second-order term—the interaction product between the two composite predictors—and only in state DS+ ($w = 6$).

Inferences on the $\Delta\mu_w$'s indicate that IDS− − has a significantly larger, and IDS++ and DS+ a significantly smaller, intercepts versus IDS−. Inferences on η_{chr}, η_{slp}, and τ_6, which are all positive and significant, suggest that the prevalence of transcription start sites increases with the chromatin structure composite predictor (and thus in particular with the number of dnase1 hypersensitive sites) in all divergence states, with the slippage composite predictor (and thus in particular with the number of microsatellite loci) in IDS−, I+, and DS+, and that these effects are further strengthened by an interaction in DS+. Complete output is provided in Table S20 in the Supplement.

Finally, to evaluate the relative contribution of each composite predictor we compute their partial R^2. On average over our 100 subsamples, these are 62.0 and 7.3%, respectively, for the chromatin structure and slippage composite predictors. The former, which is relevant in all divergence states as opposed to just three out of six, plays a much stronger role in explaining the prevalence of transcription start sites.

Figure 2 shows two 3D plots of the response against the chromatin structure and slippage composite predictors obtained through the average loadings in Figure 1. On the left are the 928 windows in state IDS−, and on the right the 464 windows in DS+. Superimposed to the points are fitted surfaces representing the final model with average coefficient estimates.

In summary, a very substantial share of the variability in transcription start sites prevalence along the human genome can be explained by using a function of the divergence states from Kuruppumullage Don et al. (2013) and as few as two composite predictors proxying chromatin structure and slippage. The former has a very strong, positive linear effect which is the same in all states. The latter is relevant only in three states, IDS−, I+, and DS+, where it has a weaker but still substantial positive linear effect. In addition, the two composite predictors have a positive interaction in DS+. The efficacy and parsimony of this exercise demonstrate that sOLS can indeed aid analysis and interpretation of high dimensional, structured genomic data.

5. Discussion

We introduced a general strategy to perform SDR when both the statistical units and the predictors in a regression are characterized by group structures. We argued, and demonstrated through simulations and a first application to genomic data, that this strategy provides an effective reduction and integration approach—with broad applicability to data that are not just high dimensional, but also complex.

Our proposal builds upon several prior developments. In particular, we combine ideas from groupwise SDR (e.g., Li et al., 2010; Guo et al., 2015) and partial SDR (e.g., Chiaromonte et al., 2002; Li et al., 2003). In principle, we could utilize any groupwise SDR method at the inner level of our procedure. However, in applications that combine high dimensions with relatively small sample sizes, it often makes sense to restrict attention to just one dominant direction for each combination of predictors' and units' groups—diminishing computational burden, and making the most parsimonious use of the limited information available. We therefore developed the novel *groupwise OLS* (gOLS) as the basis for the inner level of our *structured OLS* (sOLS). Simulation results indeed suggest that gOLS is more effective, in both computation and estimation accuracy, than other groupwise methods.

In conjunction with gOLS, we also tackled dimension estimation at the inner level. To jointly ascertain whether the relevant dimension is 0 or 1 for each predictor group (within each subpopulation), we proposed a BIC approach. Interestingly, both the simulations and our preliminary application suggest that this criterion implements a sort of "group-based" variable selection—where whole groups of predictors are eliminated from the analysis at once if they provide weak explanatory power in the context of the other groups. Intuitively, screening predictors partitioned in meaningful groups,

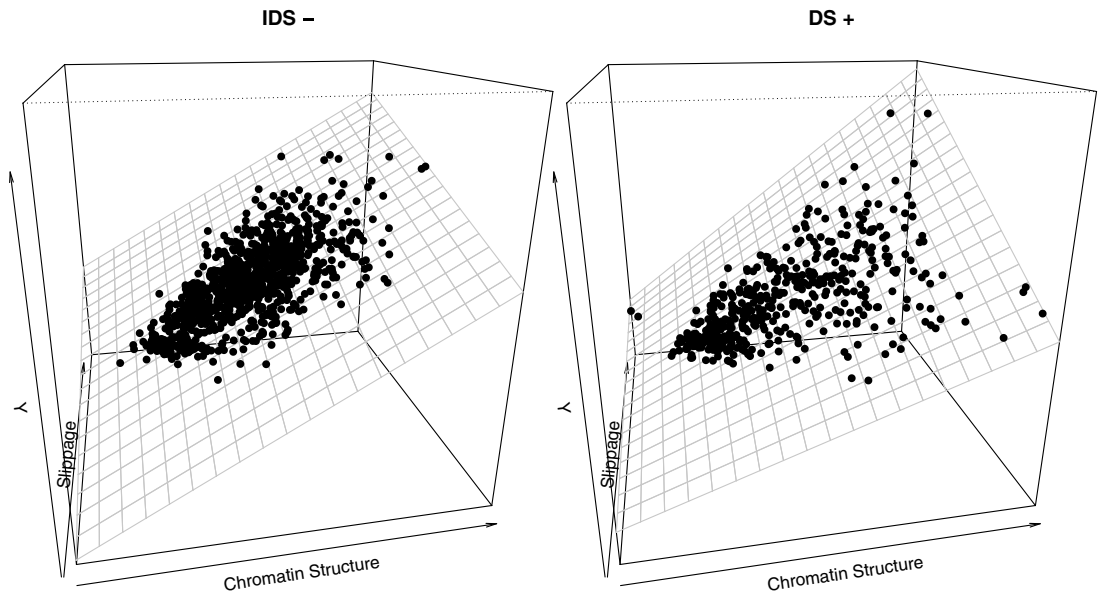

Figure 2. 3D plots of the response (coverage by transcription start sites) against the chromatin structure and slippage composite predictors for divergence state IDS− (left) and DS+ (right): all observations, that is, windows, are shown with fitted surfaces representing the final model. The equations of the surfaces are $y = 0.0085 + 0.0839x_{chr} + 0.0366x_{slp}$ (left) and $y = -0.0108 + 0.0839x_{chr} + 0.0366x_{slp} + 0.0085x_{chr}x_{slp}$ (right; with a modified intercept and an additional interaction term). The coefficients are average estimates from 100 independent random subsamples of size $n = 1000$.

as opposed to individually, could be quite efficient in some applications. We plan to investigate how this group-based variable selection compares to established methods such as lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) and elastic net (Zou and Hastie, 2005).

Notwithstanding its parsimony, sOLS still requires inversion of the sample predictor covariance matrix in each combination of predictor and unit groups—which will be singular if $n_w < p_i$ and a poor estimator even if $n_w \approx p_i$. Our current implementation can still be run when $n_w < p_i$ for some or all w's and i's, utilizing Moore–Penrose generalized inverse for the corresponding covariance matrices. Of course this could be substantially improved upon incorporating known methods for high-dimensional and under-sampled data into our structured SDR framework. For instance, shrinkage could be used to stabilize and recover invertibility of sample covariance matrices (Zhong et al., 2005), regularization/penalization techniques to sparsify OLS coefficients estimation (Li and Yin, 2008), and feature screening tools to weed out irrelevant predictors at the outset (Fan and Lv, 2008; Zhu et al., 2011). In addition, under-sampling could be tackled switching to *Partial Least Squares* (PLS) to build groupwise (gPLS) and thus structured Partial Least Squares (sPLS) along lines similar to Cook et al. (2007). This, too, will be the subject of future work.

Finally, sOLS could be generalized to yet more complex data. For example, some combinations of predictors' and units' groups may be "missing" because not all predictors are measured on units from all subpopulations, or groups may be nested or organized hierarchically. These cases are common in applications and warrant extending our methodology. Also common are multivariate responses, and we plan to incorporate multivariate SDR techniques in our framework. Hilafu and Yin (2013) already combined partial SDR (Chiaromonte et al., 2002) and projective resampling (Li et al., 2008) to deal with regressions comprising subpopulations and a multivariate responses; a similar idea could be used for sOLS.

6. Supplementary Materials

Definition S1 and S2, Lemma S1 to S3, Theorem S1 to S5, all the Proofs, as well as details of the simulation study and additional Tables and Figures referenced in Section 2.1, Section 2.2, Section 2.4, Section 3 and Section 4 are available with this article at the *Biometrics* website of the Wiley Online Library. The R package "sSDR" is publicly available on CRAN.

Acknowledgements

We thank two referees and an Associate Editor for their useful comments and suggestions, which helped us greatly in revising this work. We are grateful to K.D. Makova and our Genomics collaborators at Penn State, who provided data for the application presented in this article and created a rich and motivating context for our methodological research. In particular, we thank P. Kuruppumullage Don and R. Campos Sanchez for help with the data, and R.C. Hardison for several helpful discussions. Y. Liu and F. Chiaromonte were partially supported by funds from the Penn State Huck Institutes of the Life Sciences and the NSF. B. Li was partially supported by NSF funds.

References

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions*

of the Royal Society A, Mathematical, Physical and Engineering Sciences **367**, 4385–4405.

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics* **30**, 475–497.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics* **30**, 455–474.

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

Cook, R. D. and Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 316–342.

ENCODE Project Consortium and others (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B (Statistical Methodology* **70**, 849–911.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: Current and future challenges. *BMC Systems Biology* **8**, I1.

Guo, Z., Li, L., Lu, W., and Li, B. (2015). Groupwise dimension reduction via envelope method. *Journal of the American Statistical Association* **110**, 1515–1527.

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

Hilafu, H. and Yin, X. (2013). Sufficient dimension reduction in multivariate regressions with categorical predictors. *Computational Statistics & Data Analysis* **63**, 139–147.

Kuruppumullage Don, P., Ananda, G., Chiaromonte, F., and Makova, K. D. (2013). Segmenting the human genome based on states of neutral genetic divergence. *Proceedings of the National Academy of Sciences USA* **110**, 14699–14704.

Li, B., Cook, R. D., and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *The Annals of Statistics* **31**, 1636–1668.

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

Li, B., Wen, S., and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association* **103**, 1177–1186.

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

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

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17**, 1009–1052.

Li, L. (2009). Exploiting predictor domain information in sufficient dimension reduction. *Computational Statistics & Data Analysis* **53**, 2665–2672.

Li, L., Li, B., and Zhu, L-X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association*, **105**, 1188–1201.

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

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., and Tarczy-Hornoch, P. (2007). Data integration and genomic medicine. *Journal of Biomedical Informatics* **40**, 5–16.

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

Naik, P. A. and Tsai, C. L. (2005). Constrained inverse regression for incorporating prior information. *Journal of the American Statistical Association* **100**, 204–211.

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

Wu, S., Xu, Y., Feng, Z., Yang, X., Wang, X., and Gao, X. (2012). Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics* **13**, 320.

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

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis* **99**, 1733–1757.

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

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

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

Received September 2015. *Revised July* 2016.
Accepted July 2016.