Weng and Young *J Big Data* (2017) 4:43
https://doi.org/10.1186/s40537-017-0103-6

Journal of Big Data

SHORT REPORT Open Access

Some dimension reduction strategies for the analysis of survey data

Jiaying Weng and Derek S. Young*

*Correspondence:
derek.young@uky.edu
Department of Statistics, University of Kentucky, 725 Rose Street, Lexington, KY 40536, USA

Abstract

In the era of *big data*, researchers interested in developing statistical models are challenged with how to achieve parsimony. Usually, some sort of dimension reduction strategy is employed. Classic strategies are often in the form of traditional inference procedures, such as hypothesis testing; however, the increase in computing capabilities has led to the development of more sophisticated methods. In particular, *sufficient dimension reduction* has emerged as an area of broad and current interest. While these types of dimension reduction strategies have been employed for numerous data problems, they are scantly discussed in the context of analyzing survey data. This paper provides an overview of some classic and modern dimension reduction methods, followed by a discussion of how to use the transformed variables in the context of analyzing survey data. We highlight some of these methods with an analysis of health insurance coverage using the US Census Bureau's 2015 Planning Database.

Keywords: Big data, Central mean subspace, Flexible models, Official statistics, Principal component analysis, Sufficient dimension reduction

Introduction

The explosion of *big data* has resulted in both a dramatic increase in the volume of available data and the possibilities of how to use that data. Federal databases—which are based on survey data collected by federal agencies—are key sources of massive datasets and crucial for ongoing research. The need by researchers to analyze not only public-use data, but also restricted-use microdata, is often pivotal for addressing important research questions. The growing demand for access to such data in the United States is highlighted by the establishment of 27 Federal Statistical Research Data Centers[1], which are partnerships between federal statistical agencies and leading research institutions in the United States.

How big data can be leveraged in the construction of official statistics is a matter of ongoing discussion [1]. However, there are major benefits to how big data from federal databases, non-federal databases, or both, are used. For example, the Committee on National Statistics assembled the Panel to Review the 2010 Census. The Panel suggested more effective use of Census Bureau databases [2], which is consistent with the Census Bureau's increasing emphasis on accurate model-based predictions to conduct more efficient and cost-effective surveys [3]. Another potential benefit is to improve mutual

[1] The number of centers stated is current as of September, 2017.

government-citizen understanding [4], which in turn could improve the quality of survey data collected for federal databases.

Combining multiple federal databases, such as through record linkage techniques, can help researchers address more refined questions and produce more powerful statistical analyses. However, the resulting massive datasets often require the researcher to develop and apply a sound strategy for handling an inherently high-dimensional problem. Establishing such a strategy is also necessary for the development and dissemination of official statistics, which are based on survey data collected and stored in federal databases. Dimension reduction techniques can be an effective approach for reducing the dimensionality in big data, regardless of its source. However, there is little literature highlighting the efficacy of dimension reduction techniques in the context of analyzing survey data from federal databases. The focus of this paper fills this gap.

Lumley and Scott [5] state in the abstract of their paper, "Data from complex surveys are being used increasingly to build the same sort of explanatory and predictive models used in the rest of statistics." For example, Gelman [6] discussed the broader issue of survey weighting and regression modeling, with an application of building predictive models using the Social Indicators Survey. The analysis also developed a multilevel regression model, for which some of the standard estimation and inference procedures in those models can be applied [7]. Lumley and Scott [5, 8] demonstrated building regression models—in particular, linear and generalized linear models—using data from the National Health and Nutrition Examination Survey (NHANES). They also provided a thorough discussion about testing in such regression models being fit to survey data. Young et al. [9] demonstrated the appropriateness of using zero-inflated regression models for understanding housing unit adds or deletes in the United States based on the Census Bureau's Master Address File (MAF). In each of the above examples, inference or variable selection procedures can be employed to determine the "best" predictor variables for the respective model. However, choosing the best strategy for selecting from a large number of predictor variables can be challenging, especially in survey data where multicollinearity is almost always an issue. One appealing approach for such settings is to use dimension reduction.

We provide an analysis of data involving health insurance reform in the United States. Health insurance reform is always a major, and oftentimes controversial, social and political topic. One of the most significant efforts in recent years to health insurance reform in the United States has been the Patient Protection and Affordable Care Act (ACA), also known as "Obamacare." The ACA became effective in early 2010 with most major provisions phased in by early 2014. The ACA has an *individual mandate*, which requires each individual to buy insurance or pay a penalty if not covered by an employer-sponsored health plan or other public insurance plan. While not impacted by the provisions in the ACA, some individuals are covered under more than one health insurance plan for various reasons; e.g., supplementing coverage with a secondary plan for services not covered by their primary plan. Our example focuses on building models of health insurance coverage across the United States.

This paper is organized as follows. In "Dimension reduction techniques" section, we provide a review of principal component analysis, sufficient dimension reduction methods, and the associated algorithms. In "Flexible modeling with the transformed data"

section, we discuss how dimension reduction methods can be applied to survey data with many variables, and suggest some flexible modeling techniques that could be used with the transformed data. In "Analyzing health insurance coverage using the 2015 planning database" section, we use dimension reduction to analyze health care coverage based on survey data from the block-group-level 2015 Planning Database (PDB), which contains selected 2010 Census and selected 2009–2013 5-year American Community Survey (ACS) estimates. In "Conclusion" section, we discuss some of the conclusions from this work. Finally, in "Summary" section, we summarize what has been presented in this work as well as some general comments about dimension reduction methods.

Dimension reduction techniques

A major use of survey data is in the building of informative predictive models. For our discussion, we consider regression models with a univariate response variable Y and a p-dimensional vector of predictors $\mathbf{X}$. In full generality, the goal of regression is to characterize and infer about the conditional distribution of $Y|\mathbf{X}$. When p is large, a researcher is often faced with two major challenges, which are especially relevant to the analysis of survey data. First is that the values of the predictors are not controlled at levels as they would be in a properly designed experiment, thus, multicollinearity is often present [10]. Second is that it is often desirable to reduce the number of predictor variables, such that they are still informative about the response. These challenges can be addressed using the methods we discuss in this section. We first present principal component analysis, which is a classic and well-known multivariate procedure that can be used as a dimension reduction strategy. We then discuss more modern dimension reduction and sufficient dimension reduction techniques, including sliced inverse regression [11], partial sliced inverse regression [12], sliced average variance estimation [13], and principal Hessian direction [14, 15].

Principal components

The idea in principal component analysis (PCA) is to transform the predictor variables into linearly independent variables—or *principal components*—such that the first principal component has the largest variance, the second principal component has the second largest variance and is orthogonal to the first principal component, and so on. More formally, let Σ be the covariance matrix of $\mathbf{X}$. We want to find p linear combinations of $\mathbf{X}$ such that they are uncorrelated with each other. We construct these components such that the first component's variance is the maximum among all the linear combinations, the second component's variance is the second largest and uncorrelated to the first component, the third component's variance is the third largest and uncorrelated to both the first and the second component, etc. In other words, we wish to find the appropriate vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ such that the following holds:

- First principal component: $\text{PC}_1 = \mathbf{a}_1^{\text{T}}\mathbf{X}$, where $\mathbf{a}_1$ such that $\text{Var}(\text{PC}_1) = \mathbf{a}_1^{\text{T}}\Sigma\mathbf{a}_1 = \max_{\|a\|=1} \mathbf{a}^{\text{T}}\Sigma\mathbf{a}$;
- Second principal component: $\text{PC}_2 = \mathbf{a}_2^{\text{T}}\mathbf{X}$, where $\text{Var}(\text{PC}_2) = \mathbf{a}_2^{\text{T}}\Sigma\mathbf{a}_2 = \max_{\|a\|=1} \mathbf{a}^{\text{T}}\Sigma\mathbf{a}$ with $\mathbf{a}_1^{\text{T}}\Sigma\mathbf{a}_2 = 0$;

- pth Principal component: $\text{PC}_p = \mathbf{a}_p^{\text{T}}\mathbf{X}$, where $\text{Var}(\text{PC}_p) = \mathbf{a}_p^{\text{T}}\Sigma\mathbf{a}_p = \max_{\|a\|=1}\mathbf{a}^{\text{T}}\Sigma\mathbf{a}$ with $\mathbf{a}_p^{\text{T}}\Sigma\mathbf{a}_j = 0, j = 1,\ldots,p-1$.

The construction of the principal components are guaranteed by the following proposition, which is a condensed version of Result 8.1 in Johnson and Wichern [16]:

Proposition 1 *Let* (λ_i, η_i) *for* $i = 1, ..., p$ *be the eigenstructure of* Σ, *where* $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. *Then the* ith *principal component is given by* $\text{PC}_i = \eta_i^T\mathbf{X}$, *where* $1 \leq i \leq p$, *and* $\text{Var}(\text{PC}_i) = \eta_i^T\Sigma\eta_i = \lambda_i, \text{Cov}(\text{PC}_i, \text{PC}_k) = 0$ *for* $i \neq k$.

We can use, for example, the singular value decomposition of the covariance matrix Σ to find the eigenvalues and eigenvectors. Σ is also referred to as the *kernel matrix*. For consistency with notation used later, we denote $M_{PC} = \Sigma$ as the kernel matrix for PCA. This is estimated by $\hat{M}_{PC} = \hat{\Sigma}$, which is based on the sample data.

Principal component analysis is, perhaps, the oldest dimension reduction technique that is still widely used today [17, 18]. Consequently, PCA has been applied to numerous important data problems spanning a wide array of scientific fields. For example, PCA has been used for facial image recognition in image analysis [19], for the analysis of hormone profiles to assess the productivity of plants [20], and as part of a robust decision support tool for facilitating industrial production scheduling [21]. PCA has been applied for various survey data analyses, but due to the sometimes large number of binary or categorical variables in such data, it does not always provide reliable results [22].

Sufficient dimension reduction

Formally, a *dimension reduction* is a function $\mathcal{R}(\mathbf{X})$ that maps $\mathbf{X}$ to a k-dimensional subset of the reals such that $k < p$. Specifically, we let $\mathcal{R}(\mathbf{X}) = \mathbf{B}^{\text{T}}\mathbf{X}$, where $\mathbf{B}$ is a $p \times k$ matrix. We say that a dimension reduction is *sufficient* if the distribution of $Y|\mathcal{R}(\mathbf{X})$ is the same as that for $Y|\mathbf{X}$, which is the original conditional distribution of interest in regression models. Combining the notions of dimension reduction and sufficiency, *sufficient dimension reduction* [11, 15, 23] is used to detect a lower dimension subspace of the predictor space, such that the response variable is independent with the predictor vectors providing all the information of this subspace.

Without loss of information, $\mathbf{X}$ can be replaced by $\eta^{\text{T}}\mathbf{X}$, where $\eta \in \mathbb{R}^{p\times d}$, $d < p$. The subspace spanned by the columns of $\boldsymbol{\eta}$ is called a *dimension reduction subspace* for the regression of Y on $\mathbf{X}$. The intersection of all dimension reduction subspaces is called the *central subspace* (CS), which we denote by $\mathcal{S}_{Y|\mathbf{X}}$ with dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. The basis $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, $d < p$ of the CS has the property that $Y \perp \mathbf{X}|\beta^T\mathbf{X}$, which is to say that the conditional distribution of $Y|\mathbf{X}$ is the same as the conditional distribution of $Y|\boldsymbol{\beta}^T\mathbf{X}$. Under mild conditions [23], the CS exists and is unique.

Sometimes, the mean function $\text{E}(Y|\mathbf{X})$ may be of primary interest instead of the conditional distribution $Y|\mathbf{X}$. For such settings, Cook and Li [24] introduced the following. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, $d < p$ now be the basis for the subspace for $Y \perp \text{E}(Y|\mathbf{X})|\beta^T\mathbf{X}$. This subspace is then called the *mean dimension reduction subspace*. The intersection of all mean dimension reduction subspaces is called the *central mean subspace* (CMS), which we denote by $\mathcal{S}_{\text{E}(Y|\mathbf{X})}$.

For estimating the CS or the CMS, the following two conditions are assumed for many dimension reduction methods:

1. *Linearity condition* $\mathrm{E}(\mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X})$ is a linear function of $\mathbf{X}$.
2. *Constant variance-covariance matrix condition* $\mathrm{Var}(\mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X})$ is a non-random matrix.

In the above, $\mathbf{P}_{\mathcal{S}}$ is a projection matrix onto the subspace $\mathcal{S}$, which is either $\mathcal{S}_{Y|\mathbf{X}}$ or $\mathcal{S}_{\mathrm{E}(Y|\mathbf{X})}$ for estimating the CS or CMS, respectively.

It is important to emphasize the fundamental difference between PCA and sufficient dimension reduction. PCA reduces the number of predictors without considering the response variables, and choosing the number of principal components is not done through any formal inference paradigm. However, the idea of sufficient dimension reduction is to attain a sufficient subspace, which includes all of the information we need. There are asymptotic results for determining the number of dimensions in sufficient dimension reduction. These asymptotic results are derived and/or discussed in the references we cite for the dimension reduction techniques that we discuss below. Thus, using PCA is somewhat limited because it does not consider the response variable(s), nor does it have a formal inference mechanism for choosing the "best" number of principal components.

Sliced inverse regression

$\mathrm{E}(Y|\mathbf{X})$ is a p-dimensional surface where, for now, we assume that all of the variables represented by the columns of $\mathbf{X}$ are continuous. The notion of *inverse regression* works with the curve computed by $\mathrm{E}(\mathbf{X}|Y)$, which consists of p one-dimensional regressions. Li [11] introduced *sliced inverse regression* (SIR), which involves dividing the range of the response Y into H non-overlapping intervals called *slices*. Letting $\Sigma = \mathrm{Var}(\mathbf{X})$ and $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, we then see that $\mathcal{S}_{Y|\mathbf{X}} = \Sigma^{-1/2}\mathcal{S}_{Y|\mathbf{Z}}$ [25]. Hence, we can work on the scale of $\mathbf{Z}$. Moreover, under the linearity condition, $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|Y)} \subset \mathcal{S}_{Y|\mathbf{Z}}$ and $\mathcal{S}\{\mathrm{Var}[\mathrm{E}(\mathbf{Z}|Y)]\} = \mathcal{S}_{\mathrm{E}(\mathbf{Z}|Y)}$ [25]. Thus, we can form the kernel matrix $M_{SIR} = \mathrm{Var}[\mathrm{E}(\mathbf{Z}|Y)]$.

Let $\mathbf{x}_i$, $\mathbf{z}_i$, and y_i be the sample versions of their respective unobserved quantities. The algorithm for SIR [11] is as follows:

1. For $i = 1, \ldots, n$, standardize $\mathbf{x}_i$ into $\mathbf{z}_i$, and divide y_i into H slices. Let $\hat{f}_h$ be the proportion of the y_i in slice h, $h = 1, \ldots, H$.
2. Compute the sample mean of $\mathbf{z}$ in each slice, and denote these by $\bar{\mathbf{z}}_1, \ldots, \bar{\mathbf{z}}_H$.
3. Form the weighted variance-covariance matrix $\hat{M}_{SIR} = \sum_{h=1}^{H} \hat{f}_h \bar{\mathbf{z}}_h \bar{\mathbf{z}}_h^{\mathrm{T}}$.
4. Find the eigenstructure of $\hat{M}_{SIR}$: (λ_i, η_i), $i = 1, \ldots, p$. The $d < p$ eigenvectors corresponding to the d largest eigenvalues are the estimated directions of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|Y)}$. Then, we transform back to the original $\mathbf{X}$ scale by calculating $\hat{\beta}_l = \hat{\Sigma}^{-1/2}\hat{\eta}_l, l = 1, \ldots, d$.

Li [11] contrasted SIR with PCA by noting that the sampling properties of SIR are easy to understand and, thus, make subsequent inference using SIR fairly straightforward. SIR also is developed naturally in the regression setting, whereas PCA has to be applied

to the multivariate data consisting of only the predictors and then the response variables are regressed against the transformed predictors; i.e., this approach is called *principal components regression*. SIR has provided critical insight into various applications, such as to understand the electrochemical process of aluminum smelter plants [26] and for the purpose of direct marketing and new product design for managers of data-rich marketing environments [27].

Partial sliced inverse regression

We next consider the case where $\mathbf{X}$ can consist of both continuous and categorical predictor variables. In order to accommodate this in a setup similar to SIR, we need to use the notion of *partial dimension reduction* as introduced in Chiaramonte et al. [12]. Let W be a categorical variable with K levels and define the *partial central subspace* relative to $\mathbf{X}$ as the intersection of all subspaces spanned by $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$ such that $Y \perp \mathbf{X}|(\boldsymbol{\eta}^T\mathbf{X}, W)$. Denote the partial central subspace as $\mathcal{S}^W_{Y|\mathbf{X}}$. The relationship between partial and conditional dimension reduction is $\mathcal{S}^W_{Y|\mathbf{X}} = \bigoplus_{k=1}^K \mathcal{S}_{Y_k|\mathbf{X}_k}$, where $\mathcal{S}_{Y_k|\mathbf{X}_k}$ is the CS conditioned on level k and $\bigoplus$ is the direct sum.

For each level, the mean and covariance matrix of $\mathbf{X}_k$ are $\boldsymbol{\mu}_k$ and Σ_k, respectively. We further assume that the covariance structures are the same across the levels; i.e., $\Sigma_k = \Sigma_{pool}$, $k = 1, \ldots, K$. Now, letting $\mathbf{Z}_k = \Sigma^{-1/2}_{pool}(\mathbf{X}_k - \boldsymbol{\mu}_k)$ results in $\mathcal{S}^W_{Y|\mathbf{X}} = \Sigma^{-1/2}_{pool} \bigoplus_{k=1}^K \mathcal{S}_{Y_k|\mathbf{Z}_k}$. Then, we can use SIR for each level to find the kernel matrix M_k. After averaging these kernel matrices over different levels, we get $M^W = \sum_{k=1}^K \Pr(W = k)M_k$.

We now present the algorithm for calculating the sample version of M^W and finding the estimated directions for $\mathcal{S}^W_{Y|\mathbf{X}}$:

1. For each level k, $k = 1, \ldots, K$, calculate $\bar{\mathbf{x}}_k$ and $\hat{\Sigma}_k$, which are the sample mean and sample variance-covariance of $\mathbf{X}_k$, respectively. Moreover, calculate the common sample variance-covariance matrix $\hat{\Sigma}_{pool} = \sum_{k=1}^K \frac{n_k}{n}\hat{\Sigma}_k$ and $\mathbf{z}_{ik} = \hat{\Sigma}^{-1/2}_{pool}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)$, $i = 1, \ldots, n_k$.
2. Apply the steps in the SIR algorithm to get the sample kernel matrix in each level k: $\hat{M}_k$. Then $\hat{M}^W = \sum_{k=1}^K \frac{n_k}{n}\hat{M}_k$.
3. The first d eigenvectors, $\hat{\eta}_1, \ldots, \hat{\eta}_d$, of $\hat{M}^W$ correspond to the d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$. These eigenvectors are the estimated directions of $\mathcal{S}_{Y|\mathbf{Z}}$. Then, transform back to the original $\mathbf{X}$ scale by calculating $\hat{\beta}_l = \hat{\Sigma}^{-1/2}_{pool}\hat{\eta}_l, l = 1, \ldots, d$.

The literature on partial SIR has mostly focused on theoretical developments of the approach [12, 28, 29]. However, the efficacy of partial SIR was demonstrated in an analysis involving genomic data and clinically-relevant information in predicting survival of diffuse large-B-cell lymphoma [30]. The utility of partial SIR was also briefly highlighted in production and efficiency analyses of Norwegian electricity distribution networks [31].

Sliced average variance estimate

One disadvantage of SIR is that it cannot detect symmetric structure of predictors; however, the *sliced average variance estimate* (SAVE) method [13] can find the directions,

even in the presence of symmetric structures. Under the linearity and constant variance condition, $\mathbf{I}_p - \mathrm{Var}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Hence, $M_{SAVE} = \mathrm{E}[\mathbf{I}_p - \mathrm{Var}(\mathbf{Z}|Y)]^2$.

The algorithm for SAVE [13] is as follows:

1. For $i = 1, \ldots, n$, standardize $\mathbf{x}_i$ into $\mathbf{z}_i$ and divide y_i into H slices. Let $\hat{f}_h$ be the proportion of the y_i in slice h, $h = 1, \ldots, H$.
2. Compute the sample covariance of $\mathbf{z}$ in each slice, $\widehat{\mathrm{Var}}(\mathbf{Z}|Y = h)$.
3. Form the weighted covariance matrix $\hat{M}_{SAVE} = \sum_{h=1}^{H} \hat{f}_h[\mathbf{I}_p - \widehat{\mathrm{Var}}(\mathbf{Z}|Y = h)]^2$.
4. Find the eigenstructure of $\hat{M}_{SAVE}$ and take the first d eigenvectors, $\hat{\eta}_1, \ldots, \hat{\eta}_d$. which correspond to the d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$. These eigenvectors are the estimated directions of $\mathcal{S}_{\mathrm{Var}(\mathbf{Z}|Y)}$. Then, transform back to the original $\mathbf{X}$ scale by calculating $\hat{\beta}_l = \hat{\Sigma}^{-1/2}\hat{\eta}_l, l = 1, \ldots, d$.

SAVE has also been used for various applications and different data structures. For example, Bura and Pfeiffer [32] successfully used SAVE for class prediction of DNA microarray data. They also provided a discussion of some visualizations that can be used in such an analysis. SAVE has also been used as an effective tool in image analysis to distinguish between different facial expressions on the same person's face [33].

Principal Hessian directions

SIR and SAVE are both inverse regression approaches; i.e., we treat Y as if it were the independent variable and $\mathbf{X}$ as if it were the dependent variable. Another type of dimension reduction is the *principal Hessian direction* (PHD) method [14, 15], which is a correlation (or joint) approach. Two common types of PHDs—of which there are others—are calculated as follows:

1. Y-based PHDs: under the linearity and constant variance conditions, PHDs based on the response yield the kernel matrix $M_{yPHD} = \mathrm{E}\{[Y - \mathrm{E}(Y)]\mathbf{Z}\mathbf{Z}^T\} \subset \mathcal{S}_{Y|\mathbf{Z}}$.
2. Residual-based PHDs: PHDs based on the residuals yield the kernel matrix $M_{rPHD} = \mathrm{E}(\epsilon\mathbf{Z}\mathbf{Z}^T)$, where $\epsilon = Y - \mathrm{E}(Y) - \beta^{\mathrm{T}}\mathbf{Z}$ and $\beta = \mathrm{Cov}(\mathbf{Z}, Y)$.

The algorithm for PHD is identical to that used for SIR and SAVE, except we use the sample version of the kernel matrix corresponding to whichever of the PHDs above is of interest.

PHD has also been used for other complex data problems. For example, Cheng and Li [34] demonstrated the efficacy of using PHD in designed experiments having a large number of factors, with particular attention given to factorial designs and rotatable response surface designs. Lue [35] used PHD in the context of a regression analysis when the predictors are known to have measurement error. Lue et al. [36] showed how an imputed-spline modification to PHD yields an effective framework for conducting dimension reduction in survival regressions with censored data.

Flexible modeling with the transformed data

When building a regression model for relating a response to a large number of predictors, researchers often try fitting a multiple linear regression model first. Then, residual diagnostics are assessed to identify potential outliers, high leverage values, and overall goodness of fit. However, a multiple linear regression model is often too restrictive in practice, especially when using survey data from federal databases. Greater flexibility can be achieved using semiparametric regression models, like spline regression, generalized additive models, or partial linear models; see the texts by Ruppert et al. [37] and Härdle et al. [38] for thorough treatments of semiparametric regression modeling. The appropriateness of using such flexible models in big data settings has also been discussed in Oswald and Putka [39] and Young et al. [40].

Flexible models have been used for a wide range of analyses involving survey data from federal databases. For example, Rogers et al. [41] used cubic splines to develop migration models based on data from the ACS. Kniesner and Li [42] developed a male labor supply functions using local linear kernel regression based on panel data from the Survey of Income and Program Participation (SIPP). Gronniger [43] developed a partial linear model relating mortality to body mass index and other health measures using the data from the National Health Interview Survey (NHIS).

Each of the examples just highlighted had a large number of candidate predictor variables available from the respective survey. Many additional variables from these surveys could have been investigated by the authors for their respective model. By employing one of the dimension reduction methods discussed in "Dimension reduction techniques" section, one could develop a model of the response variable Y as a function of the d transformed predictor variables, $X_l^* = \hat{\beta}_l^{\mathrm{T}}\mathbf{X}$, $l = 1, \ldots, d$. Then, the estimated model could have better predictive ability. One example for developing such models is *principal components regression* [44], which involves estimating a multiple linear regression model for the relationship between Y and the X_l^*s, which were determined using PCA. While using a multiple linear regression model in this setup is conceptually appealing, use of visualizations may suggest the need for greater flexibility in the model. Pairwise scatterplots of Y versus each of $X_1^*, \ldots, X_d^*$ might reveal curvature or complex nonlinearities in the relationship between some of the variables, which would suggest the need for a semiparametric regression model.

The above framework is also possible when the data are from complex surveys, where population members are not sampled with equal probability. Determining appropriate survey weights is independent of the flexible modeling strategy employed with the transformed variables. Survey weights can be obtained through traditional approaches, like post-stratification and raking, or through more advanced procedures, like the flexible model-based alternatives proposed in Elliott and Little [45]. These can then be incorporated in a weighted version of the chosen semiparametric regression model, which will usually require solving a survey-weighted least squares problem [46] or implementing something like a survey-weighted backfitting algorithm [47].

Analyzing health insurance coverage using the 2015 planning database

Data

For our analysis, we use the 2015 Planning Database (PDB) [48], a publicly available Census Bureau dataset that contains housing, demographic, socioeconomic, and Census operational data. The variables and counts in the PDB are from the 2010 Census and select 5-year estimates from the 2009–2013 ACS. The data are aggregated at the block-group level. A census block is the smallest geographic unit used by the Census Bureau, and a block group comprises multiple blocks, usually containing between 600 and 3000 people. The PDB comprises approximately 220,000 block groups.

Three separate response variables are investigated for our analysis: the number of people with no health insurance coverage (Y_1), the number of people with one type of health insurance coverage (Y_2), and the number of people with two or more types of health insurance coverage (Y_3). While these could be treated as a multivariate response, we will analyze three separate models to be consistent with the dimension reduction procedures in "Dimension reduction techniques" section, which were developed assuming a univariate response. A total of 15 variables were identified as relevant candidate predictor variables. The descriptions from the PDB documentation for these variables are given in the Additional files 1, 2.

There are a total of 220,354 records in the 2015 PDB for potential analysis. We first excluded observations from the Commonwealth of Puerto Rico, which is often done due to different laws and demographic considerations involving the Commonwealth; see "Dimension reduction techniques" section of Young et al. [9] for an example of excluding Puerto Rico. The number of Puerto Rico records is 2594, which is about 1.18% of the total number 2015 PDB records. We then omitted records that had missing values for any of the variables under consideration. There are 8754 such records, which is about 3.98% of the total number of 2015 PDB records. This left us with 209,006 records for our analysis. We then transformed the predictors using the maximum likelihood approach of Box and Cox [49] in order to ensure that the linearity condition for dimension reduction is satisfied.

Analysis

The first part of our analysis focuses on reducing the dimension of our data. Each of the dimension reduction techniques discussed in "Dimension reduction techniques" section are able to be performed using functions available for the `R` programming language [50].

We first assess the presence of multicollinearity. Table 1 provides the variance inflation factors (VIFs)—a measure of the severity of multicollinearity in an ordinary least squares setting—for the 15 predictor variables. While we are not simply fitting linear models for our analysis, the use of VIFs in this context still provides a reasonable assessment of

Table 1 Variance inflation factors for the 15 predictor variables

X_1	4.7301	X_2	2.5689	X_3	7.5692	X_4	16.7708
X_5	14.5877	X_6	2.0651	X_7	6.6821	X_8	5.6688
X_9	5.7848	X_{10}	10.4067	X_{11}	3.8511	X_{12}	2.0008
X_{13}	2.1802	X_{14}	10.3657	X_{15}	11.1770		

multicollinearity. Typically, VIF values greater than 10 indicate possible influence on the least squares estimates [51, Chapter 9]. In Table 1, five variables exceed this threshold. Two of these variables—X_4 and X_5—would be expected to yield high VIFs. They are both measures of the number of ACS households with individuals who live "alone," but the former is in the context of those who live with non-relatives. While these could be essentially measuring the same effect, we will retain both of these variables for the purpose of demonstrating the efficacy of the different dimension reduction methods.

We use PCA to characterize those principal components explaining the most variation among the dataset. While Johnson and Wichern [16] state that there is "no definitive answer" to determine "how many components to retain," we proceed to use a scree plot. The scree plot consists of the principal components ordered according to their amount of variability explained on the x-axis and the cumulative proportion of the variability explained on the y-axis. The scree plot for the health insurance data is given in Fig. 1. We use 0.90 as the threshold to determine the number of principal components to select. Using this criterion, we select six principal components, which will be used for comparison with the subsequent analysis. These cumulative probabilities are also reported in Table 2.

Principal component analysis does not depend on the response variable, so the same six principal components would then be used as the predictors in the principal

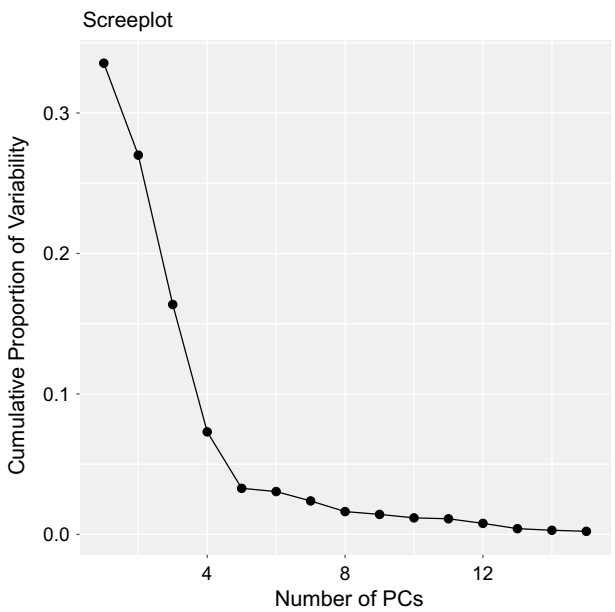

Fig. 1 Scree plot for PCA

Table 2 Cumulative proportions of variability explained using the PCA results

PC1	0.3355	PC2	0.6055	PC3	0.7691	PC4	0.8422
PC5	0.8750	PC6	0.9055	PC7	0.9294	PC8	0.9456
PC9	0.9599	PC10	0.9717	PC11	0.9828	PC12	0.9907
PC13	0.9948	PC14	0.9978	PC15	1.0000		

components regression for each of the three responses. The sufficient dimension reduction methods do, however, take into consideration the value of the independent variable. Thus, we could potentially find a different dimension for each of the three response variables.

For each of the sufficient dimension reduction procedures, testing is done to determine the dimensions. These marginal tests, based on the work in Cook [52] and Shao et al. [53], are also available in R. The tests are done sequentially, where we first test 0 dimensions versus 1 dimension, 1 dimension versus 2 dimensions, etc. Based on these tests, the dimensions selected for each of the sufficient dimension reduction procedures are summarized in Table 3. The full test results are given in the Additional files 1, 2.

SAVE, *y*PHD, and *r*PHD did not reduce the number of dimensions much or at all. Recall that partial SIR can be used when including categorical variables. A categorical variable was constructed where we partitioned the 50 states and the District of Colombia using the nine Census-designated geographical divisions [54]. The inclusion of this categorical predictor only yielded a moderate reduction according to partial SIR. The only sufficient dimension reduction procedure that noticeably reduces the dimension for each of the three responses is SIR. Therefore, the remainder of our analysis will focus on the results from PCA and SIR.

In order to use our results from PCA and SIR, we first transform the original predictor variables using the computed principal components and directional vectors, respectively. For each of the three responses, the coefficients for both methods are given in the Additional files 1, 2. We then fit an additive model [55] to each response; i.e., we fit the models

$$Y_k = \gamma_0 + \sum_{j=1}^{d_k} f_j(X_j^*) + \epsilon \tag{1}$$

for $k = 1, 2, 3$, where the f_j are unknown smooth functions of the transformed data, d_k is the dimension for the PCA or SIR results, and γ_0 is an intercept term. Thus, a total of six additive models are estimated.

For each estimated additive model, approximate t-tests can be constructed to determine the significance of each smooth term. For each of the three additive models constructed using the PCA results, all six terms are highly significant at the $\alpha = 0.05$ level. For the SIR results we found the following:

Table 3 Dimensions chosen by the marginal tests for each of the five sufficient dimension reduction methods

Method	Y_1	Y_2	Y_3
SIR	5	6	3
Partial SIR	12	12	11
SAVE	15	14	15
*y*PHD	14	15	14
*r*PHD	12	15	14

- When the response is the number of people with no insurance (Y_1), the smoothing term corresponding to the fourth dimension is not significant, with an approximate p-value of 0.102.
- When the response is the number of people with one insurance (Y_2), the smoothing term corresponding to the sixth dimension is not significant, with an approximate p-value of 0.221.
- When the response is the number of people with two or more insurances (Y_3), all of the smoothing terms are significant.

Thus, we drop the fourth and sixth terms from our models with Y_1 and Y_2 as the response, respectively, and refit the additive models. The approximate tests for all of the remaining terms yield significant results.

We also assessed the partial residual plots for each of the six fits. In the context of our additive models, the partial residual plots help us assess the relationship between the response variable and each smooth term, given that the other smooth terms are in the model. Figures 2 and 3 are the partial residual plots for the additive models based on the PCA predictors and SIR predictors, respectively, with Y_1 the response. In Fig. 2, the additive model captures some curvature for the effects due to the first four principal components (labeled as "dimensions") in this fit. In Fig. 3, the additive model captures some curvature for the effect due to the first dimension in this fit. Similar assessments can be made for the remaining four additive models. The corresponding partial residual plots are included in the Additional files 1, 2.

We next calculated the Bayesian information criterion (BIC) and adjusted R^2 values to compare the estimated additive models for each response. These results are given in Table 4. For each of the three responses, SIR yields the better BIC and adjusted R^2 values. While these measures do not provide direct comparisons between the models based on the different responses, it is worth noting that the adjusted R^2 values for the models with one insurance as a response (Y_2) are quite high relative to the other models. This indicates that there is little improvement that could be made to those estimated models by adding another set of PCA-transformed or SIR-transformed predictors.

Finally, we also assess the residuals from the additive models at the state level. Figure 4 provides maps of the United States, where the states have been shaded according to the mean of the residuals from the respective additive model built using the PCA-transformed predictors (maps in the left column) and the SIR-transformed predictors (maps in the right column). The three rows of maps correspond to those models for individuals with no insurance (Y_1), with only one insurance (Y_2), and with more than one insurance (Y_3). Notice that each pair of maps for a given response (i.e., the maps within each row) show similar distributions of the mean residuals at the state level. In particular, the maps corresponding to the additive models for Y_1 (Fig. 4a, b both show the same states with larger positive residuals, which have darker shading. These states include Nevada, Texas, Florida, and Alaska. The maps corresponding to the additive models for Y_2 (Fig. 4c, d both show that regions with larger negative residuals (lighter shading) appear mostly in the Western states while regions with larger positive residuals (darker shading) appear mostly in the Midwest. Finally, the maps corresponding to the additive models for Y_3 (Fig. 4e, f both have shading indicating residuals with overwhelmingly small magnitude.

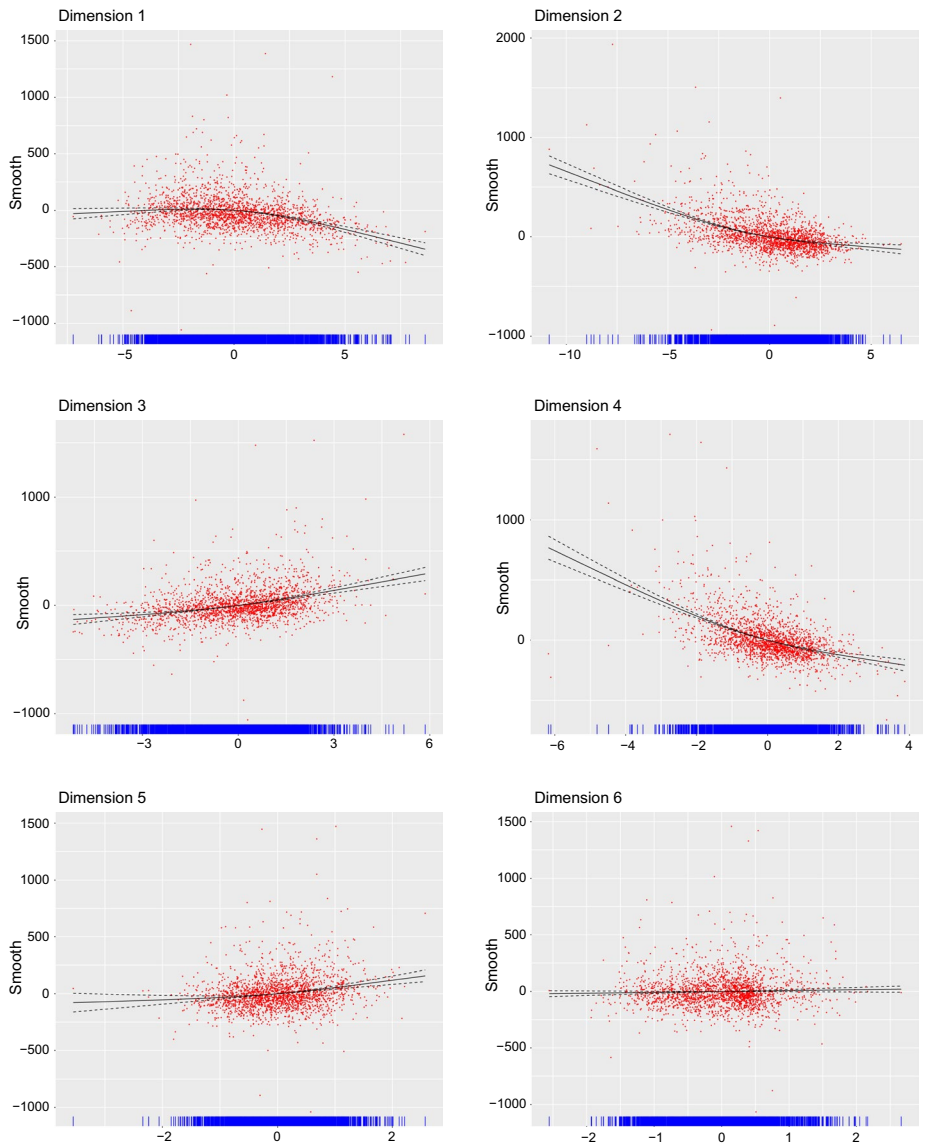

Fig. 2 Partial residual plots for each of the PCA predictors when Y_1 is the response

However, the one state indicated with a larger positive residual on both maps is Hawaii. Overall, these maps indicate that both dimension reduction strategies yield similar results for the models built for each of the three responses. Further improvements could be explored using models that, for example, include a spatial component.

Conclusion

Survey data almost always suffers from multicollinearity. When a researcher is interested in building a regression-type model using survey data, then this is bound to be an issue that they have to address. Granted this is not something unique to survey data, but it is an issue that is almost always present in survey data. Moreover, most survey datasets can be considered big data. Thus, there is a recognizable benefit to using dimension

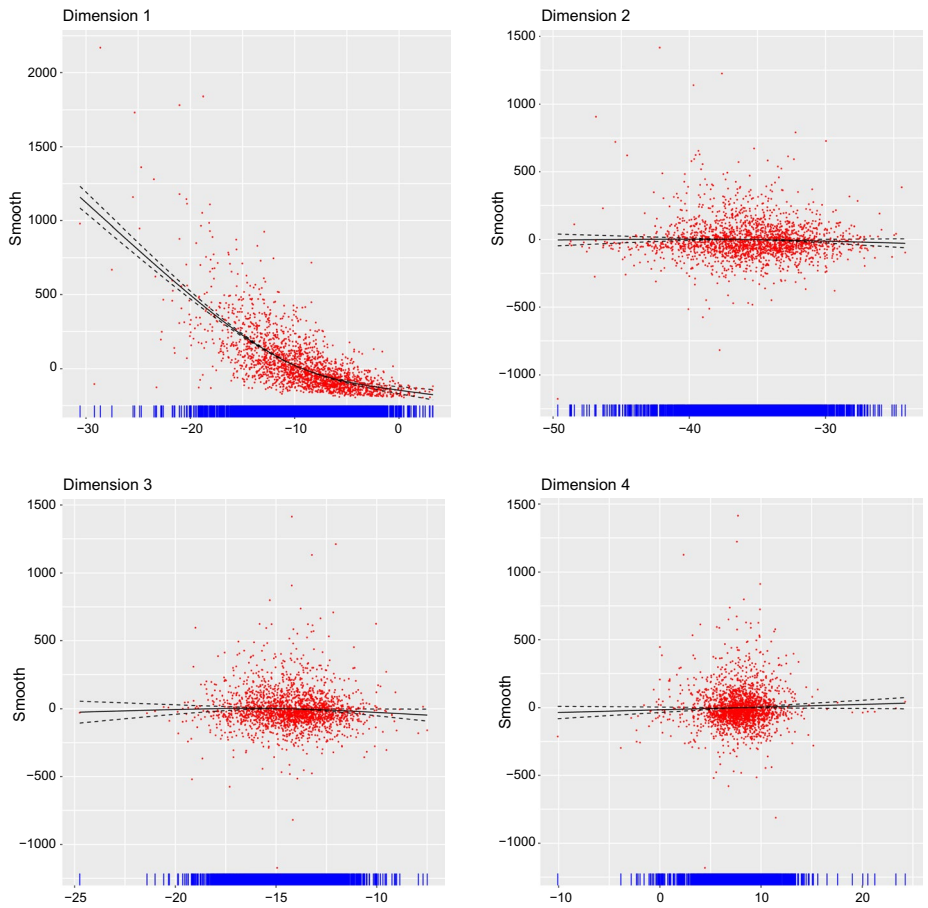

Fig. 3 Partial residual plots for each of the significant SIR predictors when Y_1 is the response

Table 4 BIC and adjusted R^2 values for each of the additive model fits using the transformed predictors from PCA and SIR

Method	Y_1	Y_2	Y_3
PCA	2695444	2906673	2564298
	0.494	0.847	0.495
SIR	2673141	2862496	2527463
	0.545	0.876	0.577

reduction techniques when building regression-type models with large survey datasets. Specifically, it can help mitigate the problems with multicollinearity as well as help reduce the dimensionality of the predictor variables under consideration.

We demonstrated the benefit of using dimension reduction procedures in the analysis of health insurance coverage data. We clearly showed that SIR provided better estimates over the other dimension reduction techniques investigated, including PCA. The other dimension reduction techniques investigated—partial SIR, SAVE, *y*PHD, and *r*PHD—did not reduce the dimensionality much for any of the three models we constructed. However, just like any other statistical analysis where you could have multiple

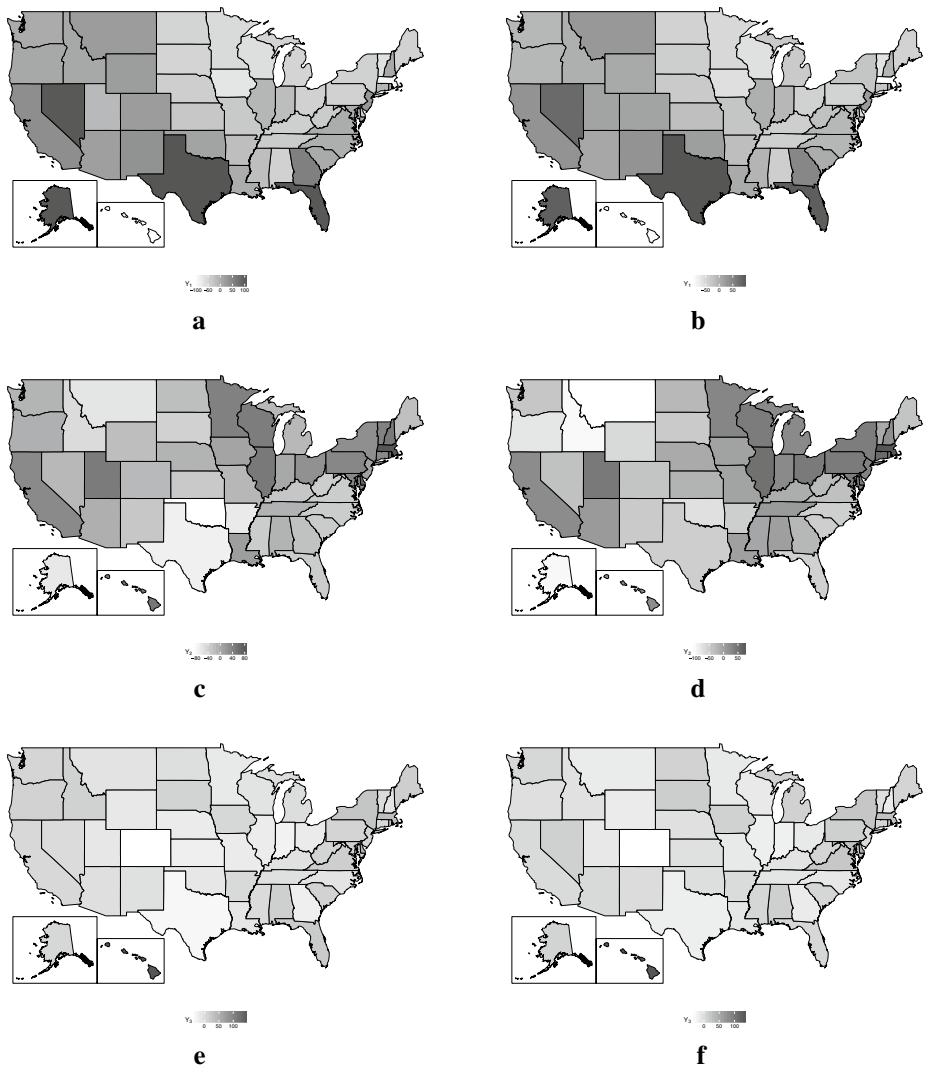

Fig. 4 Residuals from the additive models with the PCA-transformed predictors **a** Y_1, **c** Y_2, **e** Y_3 and the SIR-transformed predictors **b** Y_1, **d** Y_2, **f** Y_3

approaches to consider (e.g., different multiple comparisons procedures or different kernel methods), we advocate that the analyst consider each of the different dimension reduction procedures and then proceed to use various metrics and diagnostics to determine the best results. When taking the results from the respective dimension reduction procedure and using them in the model of interest, which for our application was an additive model, we can then use standard criteria. In our analysis, we used the BIC and adjusted R^2, both of which are well-established and accepted criteria for helping to choose between different models and assess goodness of fit. Other diagnostic plots can be constructed, such as those based on the partial residuals of the estimated model. For our application, this strategy resulted in us determining that SIR provided the best fit. From the results, we were then able to model some of the regional differences in terms of healthcare coverage.

Overall, we believe that comparing the estimated models based on different dimension reduction procedures will assist the analyst with determining the best procedure to use for their particular data problem. However, there are a few limitations that should be emphasized. One practical limitation is the availability of software. In our experience, `R` provides the most extensive collection of dimension reduction procedures available, many of which are in the `dr` package [56], but not all software have packages devoted to the implementation of dimension reduction. Another limitation is in the utility of PCA. PCA reduces the number of predictors without considering the response variable(s), and choosing the number of principal components is not done through any formal inference paradigm. Thus, the number of principal components must be chosen through a rule-of-thumb, while the same principal components must be used if building different models for multiple responses, as was the case for our health insurance analysis. Finally, the only dimension reduction technique we presented that allows for binary or categorical variables is partial SIR. Since survey data tend to have a large number of such variables (e.g., socio-economic indicators and demographic variables), partial SIR would be the only dimension reduction technique that can be directly applied to the data without requiring the analyst to do some modification to the binary/categorical variables.

Summary

Dimension reduction strategies, like PCA and sufficient dimension reduction, are being increasingly used in the era of big data. However, we believe that they are underutilized in the analysis of survey data from large databases, at least in terms of the published literature. We provided an overview of the more common dimension reduction techniques, followed by how those results can be used in flexible regression models. We then implemented that general strategy to analyze health insurance coverage data from the US Census Bureau's 2015 PDB.

The quantity of big data will continue to increase over time and this is true for data collected from large surveys. We believe that dimension reduction techniques provide an efficacious strategy for the analysis of survey data. However, it is important to acknowledge some limitations with what we have discussed in this paper.

Principal component analysis is, of course, available in most statistical software and data analytics packages. However, there is currently a limited selection of software for performing sufficient dimension reduction techniques. But as we noted in "Analysis" section, the sufficient dimension reduction techniques we employed were chosen because of their availability in `R`.

After performing dimension reduction, the resulting principal components or directional vectors help us understand features of the data that explain the most variability. However, the resulting transformed data has a more subjective interpretation. For example, in PCA, suppose demographic variables are the major contributors to the first principal component in an analysis. In this case, the analyst can attribute most of the variability in the data as being driven by demographics. But sometimes the first principal component is comprised of a subset of seemingly unrelated variables, in which case there might not be a clear interpretation.

Additional files

Additional file 1. This includes additional tables and figures for the analysis involving the 2015 Planning Database.
Additional file 2. R Code. All of the R scripts used for the analysis in "Analyzing health insurance coverage using the 2015 planningdatabase" section.

Authors' contributions
JW: Identified and prepared background material on the dimension reduction procedures discussed in the manuscript. Wrote all R scripts that were used for the analysis. Assembled summaries of all results. DSY: Identified and provided context for this applied problem. Responsible for interpreting results and identifying appropriate summaries. Responsible for preparation of manuscript. Both authors read and approved the final manuscript.

Authors' information
JW is a Ph.D. student in the Department of Statistics at the University of Kentucky. Her Ph.D. research focuses onnovel sufficient dimension reduction methods. DSY is an Assistant Professor in the Department of Statistics at theUniversity of Kentucky. His research interests include mixture modeling, tolerance regions, statistical computing, andapplied survey data analysis. Prior to joining the faculty at the University of Kentucky, he spent 3.5 years as aSenior Statistician working on data problems for the Naval Nuclear Propulsion Program and 3 years as a ResearchMathematical Statistician at the US Census Bureau working on big data problems, some of which utilized olderversions of the Planning Database. DSY is also an Accredited Professional Statistician™ of the AmericanStatistical Association.

Acknowledgements
We would like to thank Professor Xiangrong Yin of the University of Kentucky for many helpful comments on an earlier draft of this manuscript. We would also like to thank five anonymous reviewers who provided a number of important comments that helped improve the overall quality of this manuscript.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The 2015 PDB is a publicly available Census Bureau dataset located at http://goo.gl/LlcwY7. All R code used to analyze the data is available as Additional files 1, 2.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
JW was supported as a Research Assistant by NSF Grant SES-1562503 throughout the duration of this research. The funding body did not have any role in the design of the study or the collection, analysis, and interpretation of data.

Publisher's Note

Received: 23 September 2017 Accepted: 16 November 2017
Published online: 08 December 2017

References
1. Capps C, Wright T. Toward a vision: official statistics and big data. AMSTAT News. 2013;434:9–13.
2. Cook TM, Norwood JL, Cork DL. Panel to review the 2010 Census, committee on national statistics, division of behavioral and social sciences and education, national research council: change and the 2020 Census: not whether but how. Washington, D.C.: National Academies Press; 2011.
3. U.S. Census Bureau: 2020 Census operational plan: a new design for the 21st Century (2015). http://www2.census.gov/programs-surveys/ decennial/2020/program-management/planning-docs/2020-oper-plan.pdf.
4. Clarke A, Margetts H. Governments and citizens getting to know each other? Open, closed, and big data in public management reform. Policy and Internet. 2014;6(4):393–417.
5. Lumley T, Scott AJ. Fitting regression models to survey data. Stat Sci. 2017;32(2):265–78.
6. Gelman A. Struggles with survey weighting and regression modeling. Stat Sci. 2007;22(2):153–64.
7. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge: Cambridge University Press; 2007.
8. Lumley T, Scott AJ. Tests for regression models fitted to survey data. Aust NZ J Stat. 2014;56(1):1–14.

9. Young DS, Raim AM, Johnson NR. Zero-inflated modelling for characterizing coverage errors of extracts from the US Census Bureaus Master address file. J R Stat Soc Ser A. 2017;180(1):73–97.
10. Liao D, Valliant R. Variance inflation factors in the analysis of complex survey data. Survey Methodol. 2012;38(1):53–62.
11. Li K-C. Sliced inverse regression for dimension reduction (with discussion). J Am Stat Assoc. 1991;86(414):316–27.
12. Chiaramonte F, Cook RD, Li B. Sufficient dimension reduction in regressions with categorical predictors. Ann Stat. 2002;30(2):475–97.
13. Cook RD, Weisberg S. Comment on "Sliced inverse regression for dimension reduction" by Li KC. J Am Stat Assoc. 1991;86(414):328–32.
14. Li K-C. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J Am Stat Assoc. 1992;87(420):1025–39.
15. Cook RD. Principal Hessian directions revisited. J Am Stat Assoc. 1998;93(441):84–94.
16. Johnson RA, Wichern DW. Applied multivariate statistical analysis. 5th ed. Upper Saddle River: Pearson; 2002.
17. Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2(11):559–72.
18. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24(6):417–41.
19. Thomasz CE, Giraldi GA. A new ranking method for principal components analysis and its application to face image analysis. Image Vision Comput. 2010;28(6):902–13.
20. Albacete A, Ghanem ME, Dodd IC, Pérez-Alfocea F. Principal component analysis of hormone profiling data suggests an important role for cytokinins in regulating leaf growth and senescence of salinized tomato. Plant Signal Behav. 2010;5(1):45–8.
21. Mehrjoo S, Bashiri M. An application of principal component analysis and logistic regression to facilitate production scheduling decision support system: an automotive industry case. J Ind Eng Int. 2013;9(1):14.
22. Kolenikov S, Angeles G. Socioeconomic status measurement with discrete proxy variables: is principal component analysis a reliable answer? Rev Income Wealth. 2009;55(1):128–65.
23. Cook RD. Graphics for regressions with a binary response. J Am Stat Assoc. 1996;91(435):983–92.
24. Cook RD, Li B. Dimension reduction for the conditional mean in regression. Ann Stat. 2002;30(2):455–74.
25. Cook RD. Regression graphics: ideas for studying regressions through graphics. New York: John Wiley & Sons Inc; 1998.
26. Molina-Garcia A, Kessler M, Bueso MC, Fuentes JA, Gomez-Lazaro E, Faura F. Modeling aluminum smelter plants using sliced inverse regression with a view towards load flexibility. IEEE Trans Power Syst. 2011;26(1):282–93.
27. Naik PA, Hagerty MR, Tsai C-L. A new dimension reduction approach for data-rich marketing environments: sliced inverse regression. J Market Res. 2000;37(1):88–101.
28. Ni L, Cook RD. Sufficient dimension reduction in regressions across heterogeneous subpopulations. J R Stat Soc Ser B. 2006;68(1):89–107.
29. Wen X, Cook RD. Optimal sufficient dimension rreduction in regressions with categorical predictors. J Stat Plan Inference. 2007;137(6):1961–78.
30. Li L. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. Bioinformatics. 2006;22(4):466–71.
31. Orea L, Growitsch C, Jamasb T. Using supervised environmental composites in production and efficiency analyses: an application to Norwegian electricity networks. Compet Regul Netw Ind. 2015;16(3):260–87.
32. Bura E, Pfeiffer RM. Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. Bioinformatics. 2003;19(10):1252–8.
33. Ling Y. Bhandarkar SM, Yin X, Lu Q. Saveface and sirface: appearance-based recognition of faces and facial expressions. In: IEEE International Conference on Image Processing 2005, vol 2. 2005. p. 466–9.
34. Cheng C-S, Li K-C. A study of the method of principal Hessian direction for analysis of data from designed experiments. Stat Sin. 1995;5(2):617–39.
35. Lue H-H. Principal Hessian directions for regression with measurement error. Biometrika. 2004;91(2):409–23.
36. Lue H-H, Chen CH, Chang WH. Dimension reduction in survival regressions with censored dVia an imputed spline approach. Biom J. 2011;53(3):426–43.
37. Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge University Press; 2003.
38. Härdle WK, Müller M, Sperlich S, Werwatz A. Nonparametric and semiparametric models. Berlin: Springer; 2004.
39. Oswald FL, Putka DJ. Statistical methods for big data: a scenic tour. In: Tonidandel S, King E, Cortina J, editors. Big data at work: the data science revolution and organizational psychology. New York: Routledge; 2015. p. 43–63.
40. Young DS, Feng L, Charnigo RJ. Some flexible modeling paradigms for analyzing big data. J Biom Biostat. 2015;S12–e001:1–4.
41. Rogers A, Jones B, Ma W. Repairing the migration data Reported by the American community survey. Technical report, population program, Institute of Behavioral Science, University of Colorado, Boulder, Colorado; 2008.
42. Kniesner TJ, Li Q. Nonlinearity in dynamic adjustment: semiparametric estimation of panel labor supply. Empir Econ. 2002;27(1):131–48.
43. Gronniger JT. A semiparametric analysis of the relationship of body mass index to mortality. Am J Publ Health. 2006;96(1):173–8.
44. Kendall MG. A course in multivariate analysis. London: Griffin; 1957.
45. Elliott MR, Little RJA. Model-based alternatives to trimming survey weights. J Off Stat. 2000;16(3):191–209.
46. Magee L. Improving survey-weighted least squares regression. J R Stat Soc Ser B. 1998;60(1):115–26.
47. Breidt FJ, Opsomer JD, Johnson AA, Ranalli MG. Semiparametric model-assisted estimation for natural resource surveys. Survey Methodol. 2007;33(1):35–44.
48. U.S. Census Bureau: 2015 planning database (2015). https://www.census.gov/research/data/planning_database/2015/. Accessed 23 Sep 2017.
49. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B. 1964;26(2):211–52.

50. R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria; 2016. R Foundation for Statistical Computing. https://www.R-project.org/.
51. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied linear statistical models. 4th ed. New York: McGraw-Hill/Irwin; 1996.
52. Cook RD. Testing predictor contributions in sufficient dimension reduction. Ann Stat. 2004;32(3):1062–92.
53. Shao Y, Cook RD, Weisberg S. Marginal tests with sliced average variance estimation. Biometrika. 2007;94(2):285–96.
54. U.S. Census Bureau: Census Bureau regions and divisions with State FIPS Codes (2017). https://www2.census.gov/geo/docs/maps-data/maps/reg_div.txt. Accessed 23 Sep 2017.
55. Friedman JH, Stuetzle W. Projection pursuit regression. J Am Stat Assoc. 1981;76(376):817–23.
56. Weisberg S. Dimension reduction regression in R. J Stat Softw. 2002;7(1):1–22.