RedNIE | Red Nacional de Investigadores en Economía

Predictive Power of Composite Socioeconomic Indices in Regression and Classification: Principal Components and Partial Least Squares

Stefanía D'Iorio (Universidad Nacional de Entre Ríos)

Liliana Forzani (Universidad Nacional del Litoral/ CONICET)

Rodrigo García Arancibia (Universidad Nacional del Litoral/ CONICET)

Ignacio Girela (Universidad Nacional de Córdoba/ CONICET)

DOCUMENTO DE TRABAJO N° 246

Mayo de 2023

Citar como:
D'Iorio, Stefanía, Liliana Forzani, Rodrigo García Arancibia e Ignacio Girela (2023). Preferencias Parentales de Género a lo largo de tres siglos: Predictive Power of Composite Socioeconomic Indices in Regression and Classification: Principal Components and Partial Least Squares. *Documento de trabajo RedNIE N°246*.

Predictive Power of Composite Socioeconomic Indices in Regression and Classification: Principal Components and Partial Least Squares

[1], Stefanía D'Iorio[1], Liliana Forzani[2,3], Rodrigo García Arancibia[3,4], and Ignacio Girela[3,5]

[1]Facultad de Ciencias Económicas, Universidad Nacional de Entre Ríos, Entre Ríos, Argentina
[2]Facultad de Ingeniería Química - Universidad Nacional del Litoral, Argentina
[3]CONICET, Argentina
[4]Instituto de Economía Aplicada Litoral - Facultad de Ciencias Económicas - Universidad Nacional del Litoral, Argentina
[5]Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, Córdoba, Argentina

Abstract

Principal Components Analysis (`PCA`) and Partial Least Squares (`PLS`) have been used for the construction of socioeconomic status (SES) indices to use as a predictor of the well-being status in targeted programs. Generally, these indicators are constructed as a linear combination of the first component. Due to the characteristics of the socioeconomic data, different extensions of `PCA` and `PLS` for non-metric variables have been proposed for these applications. In this paper we compare the predictive performance of SES indices constructed using more than one component. Additionally, for the inclusion of non-metric variables, a variant of the normal mean coding is proposed that takes into account the multivariate nature of the variables, that we call *multivariate normal mean coding* (MNMC). Using simulations and real data, we found that `PLS` using MNMC as well as the classical dummy encoding method give the best predictive results with a more parsimonious SES index.

Keywords: Dimension Reduction, Categorical Predictors, SES, Proxy Mean Test,

1 Introduction

Composite indicators of Socioeconomic Status (SES) are a valuable tool for describing the well-being of societies, as well as for monitoring, evaluating, and implementing policies and programs aimed at improving the situation of the most economically vulnerable groups (8; 34; 12; 26). With this application in mind, during the past few decades, there has been an expansion of the use of targeted social programs for poverty reduction (40). Targeting consists of delivering services to specific groups based on a set of characteristics, namely, the poorest.

Addressing the most vulnerable group implies knowing the households' or individuals' level of well-being, generally measured in terms of income. This is the primary challenge in developing countries due to the persistence of widespread informality in labor markets. In this context, the *Proxy Means Test* (PMT) approach emerged with the aim of targeting households by calculating a score that denotes how well off the household or individual is, using a limited set of features Coady et al. (3). This score, which does not use actual income, is in most cases calculated by building a composite wealth index, also known as Socioeconomic Status (SES) index. As a result, by means of the SES index, the list of eligible beneficiaries is based on the predicted income rather than the actual income.

Selected variables for the SES index should cover a large share of the total population, be easy to measure or observe, and be difficult to manipulate by the potential beneficiaries. The most common are: geographical location of the dwelling, household quality, access to public services, demographic composition of the household, asset ownership, and labor status of the household workforce. These variables are in general non-metric variables, such as categorical, ordinal, and binary. Therefore, methods for SES index construction should consider this issue.

Once the set of variables is determined, a statistical method is chosen to predict the well-being of households. The construction of SES indices for PMT programs is mostly based on *regression* tasks, in which the level of income or consumption expenditure is predicted and then a government eligibility threshold is set (20). However, some programs may define the eligible group a priori, and a *classification* model is used to target beneficiaries instead of a regression analysis (see for example, 23).

In general, SES indices are built as a linear combination of the selected predictors variables, interpreting coefficient estimates as the associated weights to each variable. Since income varies across different geographical areas, models are often estimated separately by regions. Thus, weights differ across regions and variables are selected through an iterative procedure aim at maximizing comprehensiveness, as measured by their predictive power, i.e., how closely the estimated scores are correlated with well-being status (3).

This interest in developing weighted formulae through proxy variable selection resulted in the proliferation of dimensionality reduction techniques, primarily Principal Components Analysis (`PCA`), for the construction of the SES index (32). In fact, this approach was adopted by the World Bank and the Demographic and Health Surveys (DHS) (see, for example, 18). However, `PCA` reduces dimensionality maximizing the covariance of the variables used for the construction of the index. In this sense, Partial Least Square (`PLS`) technique has gained ground in the use of SES indices for predictive purposes since it considers the relationship between the outcome variable and predictor variables when building the weights for the composite indices (41).

Dimensionality reduction techniques (`PCA` and `PLS`) in their foundation assume variables to be metric or continuous. The theoretical relevance of having metric variables for the application of principal components can already be found in the seminal contribution of Hotelling (22), where the assumption of a definition of distance (a metric) on p dimensional space is a necessary condition for principal component application. Therefore, different methods for treating non-metric variables as continuous have emerged. First, (13) propose to create a dummy variable for each category of the non-metric variables. This approach has a number of limitations: i) it increases dimensionality, ii) it does not transform non-metric data into continuous data, iii) it misses ordinality in cases of ordinal data, and iv) it could introduce spurious correlations, since the dummy variables produced from the same ordinal variable are highly correlated (32; 27). Given these limitations, (27) propose the use of *polychoric* (for ordinal variables) and *polyserial* (for a mix of continuous and categorical variables) correlations in `PCA` computation without the need to transform the non-metric data. The idea behind these correlation estimation techniques is to assume that non-metric variables are discretized versions of underlying continuous variables, so that correlations are the maximum likelihood estimates of the correlation between a pair of unobserved normally distributed variables. In the same paper, Kolenikov and Angeles discuss a treatment method for ordinal variables that consists of estimating the mean of the marginal underlying normal variable conditional on an observed category of an ordinal variable. This procedure is

often referred to as *Normal Mean Coding* (NMC) (42). All these methods for treating the non-metric variables do not take into account the correlation relationships among variables. Forzani et al. (14), on the other hand, assume that there is a latent multivariate normal variable underlying the distribution of all data, metric and non-metric. We will explore this approach further in this paper calling it multivariate normal18 mean coding (MNMC).

Recently, there has been a growing interest in incorporating machine learning methods for SES index construction in PMT programs since it provides measures of the out-of-sample predictive power of the statistical models (29). Following the reduction approach for prediction tasks, (14) and (10) proposed supervised methodologies based on sufficient dimension reduction paradigms to build an SES index, outperforming classical non-supervised methods and other machine learning techniques (such as LASSO regression) in income prediction and poverty classification. In addition, the procedure introduced in Filmer and Pritchett (13), also called *One-Hot Encoding* (OHE) in machine learning literature, has recently regained consideration for its simplicity and has been shown to have good performance for predictive purposes (19). In fact, as shown in a comparative study by Yoon and Krivobokova (42), SES indices based on OHE treatment achieves the best results.

In some cases, for fairness purposes and civil participation, the PMT scoring system is public. Hence, households are able to check their eligibility Coady et al. (3). In addition, researchers and policymakers are interested in examining which characteristics of selected households are most closely correlated with their well-being status. Therefore, when the interpretability of the weighted formula is a matter of concern, in the applied dimensionality reduction technique of PMT models only the first component is preserved. However, even when interpretability is not required, empirical applications generally use one component, losing information and predictive power. Furthermore, as Mazziotta and Pareto (28) point out, a one-component based index frequently represents highly correlated variables while ignoring others, regardless of their potential contextual importance. Thus, the one-dimensional composite index may not capture many highly significant but poor correlation variables.

In this paper, we compare the predictive power of reduced-dimension SES indices, using `PCA` and `PLS` over a set of metric and non-metric variables using different treatment for the predictor variables. For regression problems, we test the predictive power of linear, non-linear and inverse regression. We also consider classification problems, in which we use Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Logistic Regression as predictive methods. In order to compare the treatment-reduction-prediction methodologies, we propose a group of five simulation settings based on the approach from Yoon and Krivobokova (42), and work with real data sets. Finally, in a number of empirical applications with real data, we also explore the gains of augmenting the number of component ($d > 1$) in the performance of a composite if interpretability of a SES index is not sought. To the best of our knowledge, this has not been studied in the predictive SES index literature.

In line with Yoon and Krivobokova's findings, we state the hypothesis that `PLS` with at least $d = 1$, will have higher predictive power than `PCA`. We try several prediction models for regression as well as classification. Finally, as a non-metric variables method, MNMC should produce better, or at least comparable, results than the one-hot encoding treatment method.

The rest of this paper is organized as follows. In Section 2 we describe non-metric variable treatments, dimension reduction and prediction methodologies as well as the performance metrics and validation procedures to compare the different alternatives of treatment/reduction/prediction methods. In Section 3 simulation scenarios and results from each setting are presented. In Section 4 we present different application with real data sets in regression and classification using a predictive SES index. Finally, a concluding discussion is given in Section 6. The R code we used in both simulations from Section 3 and real data analyses in Section 4 can be found at `https://github.com/stefaniadiorio/PLS-indices`. c

2 Methodology

We consider a problem where we have p predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ to construct an index to predict a response variable of interest Y (either continuous or discrete). In a given sample we have n measures of $\mathbf{X}_i$ and Y_i with $i = 1, 2, \ldots, n$ and the goal is to find a rule to predict Y in a new unobserved measure $\mathbf{X}_{\text{new}}$.

There exist a lot of supervised and non-supervised statistical methods to build composite indices when only metric variables are present. Given that in social sciences the existence of mixed variables (metric and non-metric) is predominant, some researchers have focused on how to treat them. In this paper we present novel methods to

treat mixture predictor variables and how to apply known methodologies for building a composite index.

We assume that a number q of the predictor variables are non-metric, i.e. variables measured on a non-metric scale (ordinal and nominal variables) (33). Hence, $\mathbf{X} = (\mathbf{X}_M, \mathbf{X}_{NM})$, where $\mathbf{X}_M$ and $\mathbf{X}_{NM}$ are the subset of metric and non-metric predictors, respectively. In Section 2.1 we summarize the different methodologies usually used to treat non-metric variables and a new method based on the assumption of the existence of a latent multivariate normal distribution underlying the set of predictors $\mathbf{X}_{NM}$. In Section 2.2 we present the two dimension reduction methodologies considered in this paper to build an SES index for predictive purposes, distinguishing between the unsupervised alternative (given by `PCA`) and the supervised alternative (given by `PLS`), explaining their respective algorithms. Prediction after reduction is presented in Section 2.3. After presenting the performance metrics in Section 2.4, we show how to choose the best model in simulated and a real data sets in Section 2.5. See Figure 2 for a skeleton of the methodology.

2.1 Treatment of the variables

There are several methods to treat non-metric variables in the literature, but there is no clear guidance on how to select the best one when `PCA` or `PLS` are used for prediction. In this paper we examine and apply five treatment methods. Each one entails taking into account $\mathbf{X}_M$ and possibly transforming a subset $\mathbf{X}_{NM}$ of non-metric variables. The first one (MNMC) is the new methodology proposed in the present paper, while the rest are techniques usually used in the related literature.

1. **Multivariate Normal Mean Coding** (MNMC): Based on the contribution of (15), we adopt a latent variable approach. Denote $\mathbf{Z} = E(\mathbf{V}|\mathbf{X}_{NM})$ where $\mathbf{V}$ is an underlying continuous variable of $\mathbf{X}_{NM}$ such that $(\mathbf{X}_M, \mathbf{V}) \sim \mathcal{N}(\mu, \Delta)$ and $\mathbf{V} \sim N(0, C)$. Using maximum likelihood estimators in the training data, C and Δ are estimated, and later we consider $\mathbf{X} = (\mathbf{X}_M, \mathbf{Z})$ as **metric predictors** in `PCA` and `PLS`.

2. **Normal Mean Coding** (NMC): (27): Similar to MNMC, but assuming $C = I$ and independence between $\mathbf{V}$ and $\mathbf{X}_M$. We later consider $\mathbf{X} = (\mathbf{X}_M, \mathbf{Z})$ as **metric predictors** in `PCA` and `PLS`. (27) (`PCA`) (42) (`PLS`)

3. **One Hot-Encoding** (OHE): This was presented by (13) and consists of creating a dummy variable for each category of the non-metric variables, ignoring the ordinal properties when ordinal data is present. This is called One-Hot Encoding in the machine learning literature. More precisely, consider $\mathbf{X}_{NM_j}$ to build $\mathbf{Z}$, this method transforms each non-metric variable X_{NM_j} with m_j unique categories into $m_j - 1$ binary variables. It then treats $\mathbf{X} = (\mathbf{X}_M, \mathbf{Z})$ as **metric predictors** in `PCA` and `PLS`.

4. **No Treatment-Continuous** (NTcont): In this case, no transformation of non-metric variables is performed, and they are treated as if they were metric variables. The $\mathbf{X} = (\mathbf{X}_M, \mathbf{X}_{NM})$ are considered as **metric predictors** in `PCA` and `PLS`.

5. **No Treatment-Mixed** (NTmixed): Again, no transformation of non-metric variables is performed. However, when `PCA` or `PLS` are computed, the covariance matrix used is an estimated correlation matrix. Each correlation measure between a mixture of variables is estimated by assuming an underlying bivariate normal distribution between each pair of variables. If both are continuous, the traditional Pearson correlation measure is used. When one of that pair is metric and the other one is non-metric, the *polyserial* technique is used to estimate the correlation them. If both are non-metric the *polychoric* correlation is used for pairs of polychotomous variables and *tetrachoric* correlation if both are dichotomous variables.

2.2 Dimension Reduction

Once we decide how to treat variables (Section 2.1) and before using any machine learning algorithm to predict for a new value of the predictors $\mathbf{X}$, we reduce the dimension of the predictors, making the substitution of $\mathbf{X} \in \mathbb{R}^p$ by d linear combinations of them, hopefully with $d \ll p$. We present two common approaches of dimension reduction used in composite SES index construction: Principal Component Analysis (`PCA`) as an unsupervised method (that is, it does not use information from the target variable) and Partial Least Squares (`PLS`) as a supervised method (that is, it uses in the learning stage the target variable to reduce the dimension of the predictors).

2.2.1 Unsupervised Reduction: PCA

When the goal is to decrease the variability of the prediction, researchers in social science typically use `PCA`. By preserving as much variance as possible in the predictors, `PCA` can find new uncorrelated variables that are linear combinations of the original ones that maximize variance. This technique reduces to solving an eigenvalue/eigenvector problem, i.e, if d linear combinations are used, $\mathbf{X}$ is replaced by $\mathbf{W}^T\mathbf{X}$, with $\mathbf{W}$ being the first d eigenvectors of $\mathbf{X}$'s covariance $\mathbf{\Sigma}_X$, or of $\mathbf{X}$'s correlation, $\mathbf{C}_X$, if standardized variables are used. The simplicity of `PCA` (calculating eigenvalues and eigenvectors of a matrix) makes it widely used in applied sciences. This method, as we pointed out in the introduction, is suitable for continuous variables but this is not the case of the real social science applications where nominal and ordinal variables are more common.

In this paper, we assume that we have a sample of n units from a survey and that for all of them we have measured p feature variables, that is, $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \in \mathbb{R}^{n\times p}$ with $\mathbf{X}_i^T = (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$. Depending on the way which we treat the non-metric variables (Section 2.1), we have an estimated covariance matrix $\widehat{\mathbf{\Sigma}}$ of the predictors $\mathbf{X}$. The estimator for $\mathbf{W}$ will be the largest d eigenvector of the estimated covariance matrix $\widehat{\mathbf{\Sigma}}$. Algorithm 1 shows the procedure used to compute `PCA` projections in the sample.

Algorithm 1: `PCA` (Principal Component Analysis)

1.1 Compute $\widehat{\mathbf{\Sigma}}_X = \text{Cov}(\mathbf{X})$;
1.2 Compute $\widehat{\mathbf{W}}$, the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_X$;
1.3 Define the d projections: for each data point, set i as $\widehat{\mathbf{W}}^T\mathbf{X}_i$

If not all $\mathbf{X}_i$ were in the same unit of measurement, the $\mathbf{X}_i$ may be scaled to have unit variance by taking into account $\text{diag}(\widehat{\mathbf{\Sigma}}_X)^{-1/2}\mathbf{X}_i$ rather than $\mathbf{X}_i$. The term $\text{diag}(A)$ in this context refers to the diagonal matrix, whose non-zero entries are identical to those in the diagonal of a square matrix A. In this instance, the method may be expressed as in Algorithm 2.

Algorithm 2: `PCA` (Principal Component Analysis on the correlation matrix)

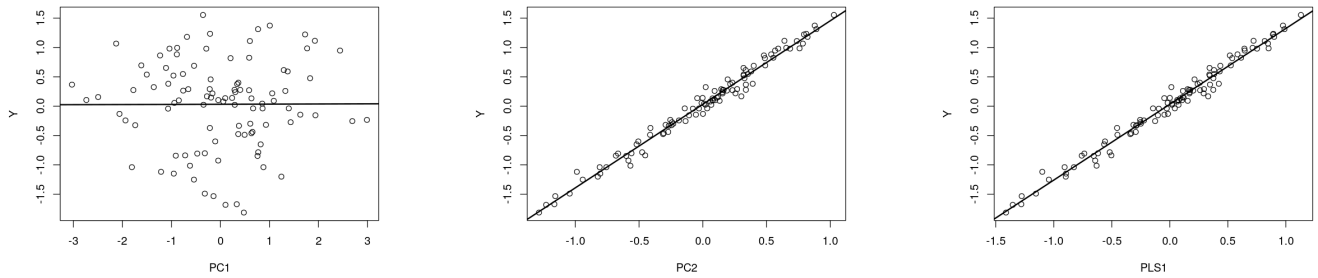
2.1 Compute $\widehat{\mathbf{\Sigma}}_{\mathbf{X}} = \text{Cov}(\mathbf{X})$;
2.2 Replace $\mathbf{X}$ by $\widetilde{\mathbf{X}} = \text{diag}(\widehat{\mathbf{\Sigma}}_X)^{-1/2}\mathbf{X}$;
2.3 Apply **Algorithm 1** to $\widetilde{\mathbf{X}}$.

The Principal Components (PCs) are orthogonal, which is an appealing feature of `PCA`'s application for regression. By using their linear combinations, all the predictors are kept in the regression function, resolving the multicollinearity issue that is frequently encountered when there are numerous predictors. It is a straightforward and appealing method, since deletion based on variance produces regression or classification function estimators with decreased variance. The issue with this argument is that, while performing regression analysis or classification tasks, we should take into consideration the prediction error which includes not only the variance but also the estimated bias (25).

The following example from (10) can help to demonstrate this point. Suppose our goal is to forecast the variable *propensity to save* (Y) from the variables *income* (X_1) and *total spending* (X_2). These three variables are related by the model $Y = X_1 - X_2 + \epsilon$, where we assume that the covariance matrix of the predictors is given by

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

which has eigenvalues $(1+r)/2$ and $(1-r)/2$ with eigenvectors $(1,1)^T$ and $(1,-1)^T$, respectively. As a result, the eigenvector corresponding to the largest eigenvalue explains $(1+r)/2\times 100\%$ of the variability, while the eigenvector corresponding to the second eigenvalue explains $(1-r)/2\times 100\%$. Assume that r is close to 1, meaning that *income* and *total spending* are highly correlated. Most of the variability in this case can be attributed to the first component. Therefore, in `PCA`, we would use $X_1 + X_2$ to predict *propensity to save* and we would discard $X_1 - X_2$, given by the second eigenvector. Clearly, this is the wrong choice. Here is the result of this example for simulated data, shown in



(a) Y vs. First PCA (b) Y vs. Second PCA (c) Y vs. First PLS

Figure 1: Simulated example where the response Y is fitted against the First and Second PCA Components and the First PLS Component

Figure 1. The data in Figure 1 is simulated using $n = 100$ observations of (X_1, X_2), which is normally distributed with $r = 0.7$. The regression model for $Y = X_1 - X_2 + \epsilon$ assumes ϵ is normally distributed with mean 0 and standard deviation 0.1. Figures 1a and 1b show the fitted response Y versus the first and second principal components as predictors (PC1 and PC2). Note how PC2 outperforms PC1 to predict Y, as suggested by their fits. Additionally, if we apply the supervised reduction method *Partial Least Square* (`PLS`) with only one component (Section 2.2.2), we observe that the results obtained with the First `PLS` projection is similar to that obtained when using the PC2, making it a better choice for the prediction index (Figure 1c).

2.2.2 Supervised Reduction: PLS

The classical Partial Least Square (`PLS`) regression algorithm estimates a linear relationship between the response Y and the predictor $\mathbf{X} \in \mathbb{R}^p$. It performs usually better than other common approaches like ordinary least squares, and it is effective for high-dimensional abundant regressions when the sample size n is insufficient in comparison to p (6; 5; 4). Cook and Forzani (7) extended `PLS` to predict Y as a function of $\mathbf{X}$ when $E(Y|\mathbf{X})$ is not necessarily a linear function of $\mathbf{X}$. This is based on a two-step algorithm. First, a dimension reduction step, where it is assumed that there are a few d linear combinations $\mathbf{W}^T\mathbf{X}$ that are sufficient for predicting the $\mathrm{E}(Y|\mathbf{X})$. Once we have the reduced predictors, these linear combinations can be used to predict Y with a general prediction rule (linear or not) if d is sufficient small. Cook and Forzani (7) proved that these linear combinations can be the first d-`PLS` projections. In the population the `PLS` reduction step is described in Algorithm 3.

Algorithm 3: `PLS` (Partial Least Square reduction in the population)

3.1 Compute $\Sigma_X = \mathrm{Cov}(\mathbf{X})$ and $\sigma_{X,Y} = \mathrm{cov}(\mathbf{X}, Y)$;
3.2 Compute $w_1 = \sigma_{X,Y}$. Set $\mathbf{W} = (w_1)$;
3.3 Compute $w_2 = Q^T_{\mathbf{W}(\Sigma_X)}\sigma_{X,Y}$. Set $\mathbf{W} = (w_1, w_2)$;
3.4 For $k = 1, \ldots, d-1$ given $w_1, \ldots, w_k$ we define $w_{k+1} = Q^T_{\mathbf{W}(\Sigma_X)}\sigma_{X,Y}$ and we set $\mathbf{W} = (w_1, \ldots, w_{k+1})$ with $Q_{v(\mathbf{\Sigma}_X)} = I - v\mathbf{\Sigma}_X(v^T\mathbf{\Sigma}_X v)^{-1}v^T\mathbf{\Sigma}_X$
3.5 Define the d projections for each data point set i as $\mathbf{W}^T\mathbf{X}_i$.

In the sample we need the estimated covariance matrix of $\mathbf{X}$ ($\widehat{\mathbf{\Sigma}}_X$), and the covariance between $\mathbf{X}$ and Y, ($\hat{\sigma}_{XY}$), defining the projection $Q_{v(\widehat{\mathbf{\Sigma}}_X)} = I - v\widehat{\mathbf{\Sigma}}_X(v^T\widehat{\mathbf{\Sigma}}_X v)^{-1}v^T\widehat{\mathbf{\Sigma}}_X$ to obtain each component iteratively. As in `PCA`, the estimaton of $\widehat{\mathbf{\Sigma}}_X$ and $\hat{\sigma}_{X,Y}$ depends on the treatment of non-metric variables.

2.2.3 From Reduction to Index: Choice of Dimensions

In the case of $d = 1$, a linear Socio-Economic Status `SES` index is defined as $\widehat{\mathbf{W}}^T\mathbf{X}$, where $\widehat{\mathbf{W}}$ is taken as weights of each variable for index construction, being attractive for economic interpretation (e.g. 15; 11). For $d > 1$, we have d

linear combinations, and therefore the traditional definition of linear SES index is insufficient. But if the goal is to obtain greater predictive power to identify individuals or households in order to implement a targeted policy, it may be convenient to use more than one dimension (i.e. $d > 1$) in the construction of the index, sacrificing interpretability of the reduction. In this case, we can define the predictive SES index as $\texttt{SES}_i = \widehat{E}(Y|\widehat{\mathbf{W}}^T\mathbf{X}_i)$. Clearly, even in the search for a predictive improvement, the lowest possible d is preferred in order to have a more parsimonious indicator. We will choose d via cross-validation using one of the performance metrics defined in Section2.4.

2.3 Prediction

Assuming $\mathbf{X}$ is a predictor, we transform it according to one of the treatment methods in Section 2.1. Afterwards, one of the two reduction methods is applied: (PCA or PLS), as described in Section 2.2, from where $\widehat{\mathbf{W}}^T\mathbf{X}$ is obtained. This section explains how to use these reduced predictors to model or predict Y, which may be continuous (regression problem) or binary (classification problem).

The regression of Y on $\widehat{\mathbf{W}}^T\mathbf{X}$ might be developed by using classical methods like linear modeling. Some settings may require only a simple transformation to induce linearity in the mean function. In addition, it is possible to use classical non-parametric regression (e.g. 17), although tuning is necessary and problems may occur if d is too large. Non-parametric regression, however, can provide a good estimator of Y if d is small, as we obtain the same asymptotic distribution as using the true reduction of $\mathbf{W}^T\mathbf{X}$ as it was proven by (Forzani et al.). When the predictors are metric, (7) presents a novel method of prediction adapted from (1). To briefly illustrate the prediction via inverse regression, let $\mathbf{Z}_d = \mathbf{W}^T\mathbf{X}$ using the population model. The paradigm behind this methodology is that it is generally easier to model adequately the inverse mean function $E(\mathbf{Z}_d \mid Y)$ than it is to model the forward mean function $E(Y \mid \mathbf{Z}_d)$, in part because $E(\mathbf{Z}_d \mid Y)$ can be visualized straightforwardly in d scatter plots, one for each component of $\mathbf{Z}_d$ vs Y. We present in what follows a method for estimating the forward mean function $E(Y \mid \mathbf{Z}_d)$, without explicitly specifying a model for the forward regression, using the inverse regression of $\mathbf{Z}_d$ on Y. The conditional density of $\mathbf{Z}_d \mid Y$ is denoted by $g(\mathbf{z}_d \mid Y)$. Then for Y continuous or binary,

$$E(Y \mid \mathbf{Z}_d = \mathbf{z}_d) = \frac{E\{Y g(\mathbf{z}_d \mid Y)\}}{E\{g(\mathbf{z}_d \mid Y)\}}, \tag{1}$$

where all right-hand side expectations are with respect to the marginal distribution of Y. A sample version of (1) provides an estimator of the forward mean function:

$$\begin{aligned} \widehat{E}(Y \mid \mathbf{Z}_d = \mathbf{z}_d) &= \sum_{i=1}^{n} \xi_i(\mathbf{z}_d) Y_i \\ \xi_i(\mathbf{z}_d) &= \frac{\widehat{g}(\mathbf{z}_d \mid Y_i)}{\sum_{i=1}^{n} \widehat{g}(\mathbf{z}_d \mid Y_i)}, \end{aligned} \tag{2}$$

where $\widehat{g}$ denotes an estimated density. This estimator is quite different from the standard non-parametric kernel estimator (eg. 36, Ch. 4). As opposed to kernel estimators, here the weights ξ_i depend on Y, and as our weights are determined exclusively from $\widehat{g}$, bandwidth estimation is not necessary.

Multivariate normal models are used to estimate densities

$$\mathbf{Z}_d = \mu_d + \Theta_d f(Y) + \varepsilon_d, \tag{3}$$

where $f \in \mathbb{R}^s$ is a known user-specified vector-valued function of the response that does not depend on d, Θ_d is a $d \times s$ matrix of regression coefficients and $\varepsilon_d \sim N_d(0, \Sigma_{Z_d|Y})$ is independent of Y. This model is in line with the chemometrics models proposed by (30).

In this paper, after obtaining the PCA or PLS projections, we will use the following methodos for prediction in regression: Linear Regression (LM), Non-parametric Regression (NP) and Inverse Regression (IR). For classification we use Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Inverse Regression (IR). (24; 21)

2.4 Performance Metrics

Since the goal of this paper is to predict a response variable Y using linear combinations of metric and non-metric variables $\mathbf{X}$, the assessment of the achieved performance needs to be driven by prediction accuracy. For continuous response, once we estimate `PCA` or `PLS` projections and perform predictions using linear, non-parametric, or inverse regression, the error measure is computed via the mean-squared error (MSE). Following Duarte et al. (11), for classification problems, when Y is binary, we use the Area Under the Curve (AUC), which gives a performance measure of the classification method trade-off between sensitivity (TPR) and specificity (TNR). We also use the misclassification error (Misclas) and the Index of Union (IU) defined (see 38) as

$$IU(c) = (|TPR(c) - AUC|) + (|TNR(c) - AUC|) \tag{4}$$

where c is the cut-off threshold c^* and $TPR(c)$ is the True Positive Rate and $TNR(c)$ is the True Negative Rate, both as function of c. We include TNR and TPR metrics since $1 - TPR$ and $1 - TNR$ are also measures of the *undercoverage* and *leakage rates* of targeted policy programs. The former is the proportion of poor households that are not included in the program (errors of exclusion) whereas the latter is the proportion of those who are reached by the program who are classified as non-poor (errors of inclusion) (3).

2.5 Validation

Each treatment-reduction-prediction methodological alternative will be a proposed method, and we compare them using the k-fold cross-validation procedure. That is, first we randomly divide the sample into k groups (folds) of approximately equal size. One of the folds is used as a test set and the rest $k - 1$ folds are used as a training set. In the training set, all the treatment techniques presented in Section 2.1 are applied, `PCA` and `PLS` projections are computed for all $d = 1, \ldots, \min\{p, n\}$ and regression or classification models are trained. We used all the different (treatment-reduction and regression or classification) models to predict in the test data set. Using the performance metrics defined in Section 2.4 we get the errors in the test data set. This procedure is repeated k times for each fold, obtaining k error measures, reporting an average of the errors. The optimal dimensions d^* for `PCA` and `PLS` reductions are those that minimize the cross-validation MSE in regression problems and maximizes AUC in classification.

We later refit the best model using all the data, and we use this model to predict a new $\mathbf{X}_{\text{new}}$.

3 Simulation

We define a series of settings to study how the different proposed methods work. We assume a multivariate normal joint distribution for all the variables in $\mathbf{X}$, where a proportion of them are observed (the metric variables) and another are an unobserved set of latent variables. The subset of non-metric variables is generated from these latent variables. The approach of latent variables models assumes that there are no directly observable variables, albeit manifested in a set of other observable variables (2). We first simulate a vector $\mathbf{X} \in \mathbb{R}^p$ of metric variables (some of them will be used as latent variables to generate the non-metric ones) and, subsequently, the outcome continuous variable Y as a function of $\mathbf{X}$. We also consider a classification problem where in all the settings Y was converted to binary data with value 1 if the simulated Y was greater than the median of the simulated Y and 0 otherwise.

Secondly, since ordinal variables are the non-metric variables that predominate in socioeconomic data, a proportion p_{nm} of $\mathbf{X}$ is discretized following this nature. That is, for the discretized ones, the i-th observation of the j-th variable of $\mathbf{X}$ are discretized in m_j unique categories according to the function:

$$x_{ij} = \begin{cases} 0, & \text{if } x_{ij} \leq \tau_{j,1} \\ 1, & \text{if } \tau_{j,1} < x_{ij} \leq \tau_{j,2} \\ \vdots & \vdots \\ m_j - 2, & \text{if } \tau_{j,m_j-2} < x_{ij} \leq \tau_{j,m_j-1} \\ m_j - 1, & \text{if } x_{ij} > \tau_{j,m_j-1} \end{cases}$$

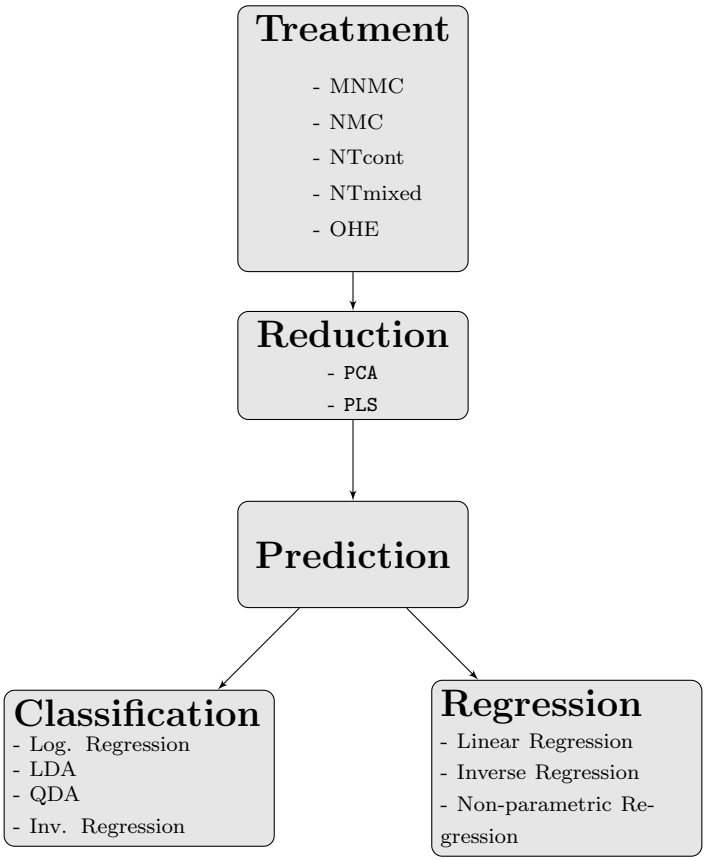


Figure 2: Simplified flow diagram of every treatment-reduction-prediction method used in this paper

where $\tau_j = \{\tau_{j,1}, \ldots, \tau_{j,m_j-2}\}$ are the thresholds for the j-th variable which are generated as

$$\tau_j = \left(F^{-1}(u_{j,1}), \ldots, F^{-1}(u_{j,m_j-1})\right)$$

where $F(\cdot)$ is the empirical CDF of the j-th variable realizations and $\left(u_{j,1}, \ldots, u_{j,m_j-1}\right)$ are generated randomly from a uniform distribution on $[0, 1]$ and sorted in ascending order.

In all cases, we use $p = 50$ predictor variables and two training sample sizes: $n = 100$ and $n = 1000$ and a test sample of $n = 500$ to evaluate the performance of the selected models with the training set. The proportion p_{nm} of the non-metric is 80% for **S1**, **S2**, **S3** and **S4** and 30% for **S5**. We consider d, i.e., the number of components needed for `PLS` and `PCA`, as known, but in some cases we estimate the parameters using only one dimension to show the effect of underestimating the number of projections used. This experiment is repeated 100 times and results are reported in a boxplot of the performance metrics in the test data defined in Section 2.4.

3.1 Settings

We generate $\mathbf{X}$ in five different ways: **S1**-**S5**. Settings **S1** and **S2** were studied by (42) and are the benchmark for the subsequent settings. We build **S3** upon **S2** assuming a non-linear model between the response and the predictors to show the behavior of the `PCA` and `PLS` under non-linearity. This nonlinear behavior is revealed in several investigations that contemplate the association between the SES index and some response variable of interest (e.g. 37; 35; 9). Configuration **S4** is a modification of **S2** in the sense than now Y depends on two linear combinations of `PCA` and `PLS` instead of one. This is intended to show the effect on the predictive performance of the missing information when a dimension of 1 is used to estimate instead of 2. We include this scenario since the possibility of having more than one component to represent an SES index is totally plausible as in (39). For example, (31) define the SES taking the first PC as "material-based" SES index and the second PC for a "social-based" SES index. Finally, setting **S5** seeks to show the fact that `PCA` is unsupervised and so it can miss important information when we consider only few components, while `PLS`, being supervised, can get them. As shown in (10), the difference between a supervised SES index and an unsupervised one can generate a significant gap in the sought predictions. We generated the metric variables and the continuous response Y as:

- **S1: A single latent factor related with Y and $\mathbf{X}$**

  $$\mathbf{X} = X_1\lambda_1 + \Delta$$

  where $\lambda_1 = \frac{1}{\sqrt{p}}1_p$ represents the loading, 1_p is a row vector $1 \times p$ of ones, $X_1 \sim N(0, 1)$ is the common latent factor and $\Delta \sim N_p(0_p, (9p)^{-1}\mathbf{I}_p)$. For the regression model, $Y = \beta_1 X_1 + \epsilon$ with $\beta_1 = 1$ and $\epsilon \sim N(0, 0.01)$ independent of X. In this case $\dim(\texttt{PCA}) = 1$ and $\dim(\texttt{PLS}) = 1$ with $\mathbf{W}_{\texttt{PLS}} = \mathbf{W}_{\texttt{PCA}} = (1, 0, \ldots, 0)^T$. For estimation and prediction we used a dimension of 1.

- **S2: A latent factor related with Y and $\mathbf{X}$, and another orthogonal factor related with X**

  $$\mathbf{X} = 3.5X_1\lambda_1 + 0.3X_2\lambda_2 + \Delta$$

  where again $\lambda_1 = \frac{1}{\sqrt{p}}1_p$ and λ_2 is the vector $(\frac{1}{2\sqrt{p}}\mathbf{1}_4^T, \frac{-2}{\sqrt{p}})$ repeated $\frac{p}{5}$ times. Note that $\|\lambda_1\| = \|\lambda_2\| = 1$ and $\lambda_1 \perp \lambda_2$. $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 5)$ represent two independent common latent factors, and $\Delta \sim N_p(0_p, (9p)^{-1}\mathbf{I}_p)$. $Y = \beta_1 X_1 + \epsilon$, where $\beta_1 = 1$ and the error term $\epsilon \sim N(0, 0.01)$. The first latent factor X_1 is related to the response variable, but X_2 doesn't have any influence on it, so in this case $\dim(\texttt{PCA}) = \dim(\texttt{PLS}) = 1$ with $\mathbf{W}_{\texttt{PLS}} = \mathbf{W}_{\texttt{PCA}} = (1, 0, \ldots, 0)^T$. For estimation and prediction we used a dimension of 1.

- **S3: A latent factor related with Y and $\mathbf{X}$, and another orthogonal factor related with $\mathbf{X}$** and Y is non-linear on $\mathbf{X}$.

  This setting has the same parameters as **S2**, but after we simulate Y we consider Y^2 as the response, and therefore $\dim(\texttt{PCA}) = \dim(\texttt{PLS}) = 1$ with $\mathbf{W}_{\texttt{PLS}} = \mathbf{W}_{\texttt{PCA}} = (1, 0, \ldots, 0)^T$. For estimation and prediction we used a dimension of 1.

- **S4: A latent factor related with Y and X, and another non-orthogonal factor related with X**

$$\mathbf{X} = 3.5X_1\lambda_1 + 0.3X_2\lambda_2 + \Delta$$

where again $\lambda_1 = \frac{1}{\sqrt{p}}1_p$ and $\lambda_2 = 4(\frac{\sqrt{2}}{\sqrt{p}}\mathbf{1}_{\frac{p}{2}}^T, 8\mathbf{1}_{\frac{p}{2}}^T)$. Note that here $\lambda_1 \not\perp \lambda_2$. $X_1 \sim N(0,1)$ and $X_2 \sim N(0,5)$ represents two independent common latent factors, and the error term $\Delta \sim N_p(0_p, (9p)^{-1}\mathbf{I}_p)$. Only the first latent factor X_1 is related to Y as $Y = \beta_1 X_1 + \epsilon$ where $\beta_1 = 1$ and the error term $\epsilon \sim N(0, 0.01)$. Nevertheless $\dim(\texttt{PCA}) = \dim(\texttt{PLS}) = 2$ with $\mathbf{W}_{\texttt{PLS}} = \mathbf{W}_{\texttt{PCA}}$ are the largest two eigenvectors of the matrix $(3.5)^2\lambda_1\lambda_1^T + (.3)^2 25\lambda_2\lambda_2^T$. We estimate and predict using the true dimension 2 and dimension 1 to show the loss of information if we underestimate the dimension.

- **S5: Forward setting**

$$\mathbf{X} \sim N(0, \mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is a simulated correlation matrix that we fix at the beginning of the simulation and is the same for each repetition. Once we generated $\mathbf{\Sigma}$ we define $W_{\texttt{PLS}} = (v_7, v_8, v_{13})$ where v_i indicates the i eigenvector of Σ. Once $W_{\texttt{PLS}}$ is defined, $Y = v_7^T\mathbf{X} + v_8^T\mathbf{X} + v_{13}^T\mathbf{X} + \epsilon$ with $\epsilon \sim N(0, 0.25)$. Therefore, in this case $\dim(\texttt{PCA})$=13 and $\dim(\texttt{PLS})$=3. We estimate and predict using dimension 3 for PCA and PLS, so that we are underestimating the dimension needed for PCA.

3.2 Simulation Results

The results of the simulation (performance metrics) are presented in boxplots in Appendix A, one for data generating processes, faceted by each variable treatment, reduction and prediction methodologies presented in Section 2. Specifically, we present the boxplots for $n = 100$ and only a few ones for the case $n = 1000$, since when the true number of directions is used the only difference is that $n = 100$ gives wider boxplots, indicating that the methods have more variability for smaller n. When the number of components used is not the true one (like in the case of **S4** and **S5** for PCA) not only is the variability for $n = 100$ larger than that for $n = 1000$, but the precision is smaller, increasing the average MSE as can be seen, for example, when comparing Figure 6 with Figure 7.

In general, for the regression problems, the best methods for non-metric variable treatment were MNMC and OHE. For MNMC, this result is expected since the variables used in the end contain the most information about their joint distribution. In fact, the data is generated as a multivariate Gaussian distribution, and therefore MNMC would be the most appropriate method. However, the results show that the inclusion of dummies through the OHE is also a good strategy. On the other hand, NM coding omits the correlation among the predictors and that translates to a loss of efficiency in the prediction. The same is true if we do not treat the variables, since we lose their nature and that translates into poorer predictive power. In **S5**, MNMC-PLS and NTmixed-PLS display similar performances, with lower MSE, although MNMC presents more variability in non-parametric regression.

In settings **S1**-**S4**, the theoretical PCA and PLS are the same and, as expected, when using the true number of directions the performance of PCA and PLS reduction is similar. Nevertheless, in **S4**, where we use only one projection, PLS gives better results than PCA. This is not surprising since PLS tries to capture in one direction as much information about Y as possible while PCA is unsupervised. Also, OHE-PLS is clearly the best in **S4** considering one component, for both sample sizes. Concerning **S5**, as expected, PCA gives much worse results than PLS since for PCA, the data is projected into the first three eigenvectors and the true ones are eigenvectors 7, 8 and 13 while the first 3 PLS capture all the information from those components.

It is not surprising that in settings **S1**, **S2**, **S4** and **S5**, linear prediction gives the best results since Y depends linearly on the PCA and PLS projections. Nevertheless, not much efficiency is lost if inverse regression or non-parametric prediction is used on those settings. But when non-linearity is present, as in **S3**, the performance of the linear model drops. Then, if we are not sure that the model is linear, the non-parametric and inverse regression approach will be more adequate. The computational cost should be taken into account in the decision process, since non-parametric and inverse regressions are computationally demanding.

In classification, MNMC is overall the best treatment methodology. In relation to the reduction methods, we reach the same conclusion as in the regression problem. Logistic regression and LDA give very similar results, as it happens in general. We omitted quadratic discriminant (QDA) since, in the proposed settings, it never gives better results. This is due to the fact that the simulated models have constant variable and therefore LDA applies. QDA applies too, but since it requires the estimation of more parameters the results are not as good as when we apply LDA.

4 Experiments with Real Data

In order to compare the performance of the treatment-reduction-prediction methodologies (hereafter we refer a treatment-reduction-prediction method simply as model) in regression (continuous response) and classification (binary response) using real data, we use two different data sets: the Argentinian Household Survey (*Encuesta Permanente de Hogares-EPH*) for the second quarter of 2017 and the Argentinian National Household Expenditure Survey of 2018 (*Encuesta Nacional de Gastos de los Hogares-ENGHO*). These surveys are carried out by the *Instituto Nacional de Estadística y Censos* from Argentina (INDEC) and we restricted our attention at the household level for the Greater Buenos Aires (GBA) demographic area, where more than 30 percent of the Argentinian population is concentrated.

The application for regression and classification are summarized as following:

- Regression
  - Response: Logarithm of the household total expenditure.
    Dataset: ENGHO - GBA.
    Number of variables (metric/non-metric): 20 (5/15).
- Classification:
  - Response: Poor or non-poor households in monetary terms (if the household adult equivalent income is lower than the official poverty line).
    Dataset: EPH - GBA.
    Number of observations: 2329.
    Number of variables (metric/non-metric): 26 (5/21).

A summary of the predictor variables (metric and non-metric) for each dataset is provided in Table 1. The number of categories for each non-metric variables is specified in brackets. Note that in both cases the response variables are proxy measures of household well-being status. This is because we are examining the predictive power of a composite SES index to be used in targeted programs, specially in a Proxy Means Tests scheme.

Targeted social programs usually rely upon external data sources such as official household surveys representative of the population of interest. In other cases, an specific household survey may be designed for collecting data for the program since external household surveys are usually topic-specific (such as labour force) or cover specific population subgroups (children and youth) that may be different from what is intended by the program. Hence, the sample of the survey used to train the model may be reduced substantially either when the program is restricted to a smaller geographical area or specific subgroup.

For this latter reason, we first train the models with a much lower sample size than the complete data set (such as $n = 200, 250, 400, 600$) with a 5-fold Cross Validation for selecting the optimal model using the performance metrics defined in Section 2.4. We report the final 5-fold cross validation (CV) errors.

5 Results for Real Data

5.1 Performance in Regression: Prediction of Total Expenditure

Tables 2 and 3 show the predictive performance for every model in the context of regression for a training sample of $n = 600$ and $n = 200$ respectively. In the first column (Dimension) the optimal d^* that minimizes the MSE in the

Data set	**EPH**	**ENGHO**
Variables		
Metric		
	# Active members	Household size
	# Older adults	# Older adults
	# Children	# children (under 14)
	Overcrowding	Overcrowding
	# working hours of hh	Dependency ratio
Non-metric		
	Source of drinking water (4 cat.)	Source of drinking water (4 cat.)
	Housing quality (3 cat.)	Housing quality (3 cat.)
	Toiled drainage (3 cat.)	Toilet drainage (3 cat.)
	Toilet facility (3 cat.)	Toilet facility (2 cat.)
	Toilet sharing (3 cat.)	Has car? (3 cat.)
	Water location (3 cat.)	Housing tenency (3 cat.)
	Toilet location (2 cat.)	Has garden? (3 cat.)
	Highest education level of hh (7 cat.)	Has air acond.? (2 cat.)
	Lives near to landfill (2 cat.)	Medical coverage (2 cat.)
	Lives in a floodplain (2 cat.)	Disabilities (3 cat)
	Lives in a shanty town (2 cat.)	Cooking fuel (2 cat.)
	Cooking fuel (2 cat.)	Educational climate (5 cat.)
	Has separate Kitchen ?(2 cat.)	Highest education level of hh (7 cat.)
	Has Laundry? (2 cat.)	Gender of hh (2 cat.)
	Has garage? (2 cat.)	Hh currently working? (2 cat.)
	Receive monetary aids? (2 cat.)	
	Receive goos aids? (2 cat.)	
	Receive other aids? (2 cat.)	
	Gender of hh (2 cat.)	
	Medical coverage (2 cat.)	
	Hh currently working? (2 cat.)	

Table 1: Variables used to compute a composite SES index to predict or classify different response variables

5-fold CV experiment for $d = 1, \ldots, p$ is reported. In the MSE columns we report the Mean Square Errors of the 5-fold CV obtaining the d^* for each treatment, reduction and predictive method. The last column shows the MSE for $d = 1$.

The last two columns can be used to compare `PCA` and `PLS`-based models in context where an interpretive predictive SES index is of interest ($d = 1$) and the gains in prediction performance from increasing the number of components ($d^* > 1$).

Following the skeleton of this treatment-reduction-prediction methodology in Figure 2, for regression problems, we have 40 models to compare overall. Therefore, bearing in mind the formulated hypothesis in Section 1, we describe the patterns of interest.

The results are explained in the following way: First, we compare performances when $d = 1$ and later we analyze the benefits of increasing d in terms of predictive power at the risk of sacrificing interpretability.

For traditional SES index construction with $d = 1$, `PLS` yields better results than `PCA` regardless of which treatment method is used with the exception of NMC and NTmixed. However, these two treatment methods are not able to outweigh NTcont-based models where non-metric variables are not treated. The best models are those based on MNMC and OHE as treatment and `PLS` as reduction. The rest of the models are not even superior to our baseline model: NTcont.

Considering the possibility of increasing the number of components to augment the predictive power, in general terms, the performance of every model is better with d^* than $d = 1$. `PCA`-based models benefit the most from increasing the number of components. In models which use OHE and MNMC as treatment, `PCA` yields comparable results to `PLS`. This was not the case when $d = 1$. Finally, `PCA` needs more components than `PLS` to achieve the lowest MSE. Hence, `PLS` results in more parsimonious models than `PCA`. Models based on OHE and MNMC as

treatment are strictly superior to the rest.

If we reduce the number of observations in the sample ($n = 200$) some new patterns emerge (Table 3). First of all, the overall performance of the model is poorer than with larger data samples ($n = 600$). Furthermore, when $d = 1$, the best models are those based on MNMC and OHE as treatment and `PLS` as reduction and, among these, there are no substantial differences between the prediction methods. The rest of the models based on NMC and NTmixed don't give better performance than NTcont. Not even when we search for the optimal number of components d^*. Notwithstanding, this is not the case for models based on MNMC and OHE as treatment but based on `PCA` as reduction. As is the case with larger sample size, we observe that `PCA` benefits more than `PLS` from using the optimal d and results between both reduction methods are now comparable. However, again, `PCA` requires a larger number of components to achieve at least as good predictive power as `PLS`.

d^*	Treatment	Reduction	Model	$MSE(d^*)$	$MSE(d=1)$
12	MNMC	PCA	Inverse.Reg	0.3721	0.5969
15	MNMC	PLS	Inverse.Reg	0.3504	0.3938
15	MNMC	PCA	Linear.Reg	0.3707	0.5931
13	**MNMC**	**PLS**	**Linear.Reg**	**0.3486**	**0.3943**
15	MNMC	PCA	Non.Parametric.Reg	0.3980	0.6098
7	**MNMC**	**PLS**	**Non.Parametric.Reg**	**0.3821**	**0.3940**
11	NM	PCA	Inverse.Reg	0.6340	0.6407
2	NM	PLS	Inverse.Reg	0.7611	0.8058
1	NM	PCA	Linear.Reg	0.6407	**0.6407**
2	NM	PLS	Linear.Reg	0.8041	0.8133
2	**NM**	**PCA**	**Non.Parametric.Reg**	**0.6335**	0.6631
2	NM	PLS	Non.Parametric.Reg	0.7279	0.7837
11	NT cont	PCA	Inverse.Reg	0.4820	0.4868
1	NT cont	PLS	Inverse.Reg	0.4726	0.4726
1	NT cont	PCA	Linear.Reg	0.4866	0.4866
1	**NT cont**	**PLS**	**Linear.Reg**	**0.4715**	**0.4715**
1	NT cont	PCA	Non.Parametric.Reg	0.4925	0.4925
1	NT cont	PLS	Non.Parametric.Reg	0.4959	0.4959
4	**NT mixed**	**PCA**	**Inverse.Reg**	**0.4908**	**0.5717**
9	NT mixed	PLS	Inverse.Reg	0.4992	0.5987
4	NT mixed	PCA	Linear.Reg	0.5064	0.5697
9	NT mixed	PLS	Linear.Reg	0.5201	0.5917
3	**NT mixed**	**PCA**	**Non.Parametric.Reg**	**0.4996**	**0.5704**
9	NT mixed	PLS	Non.Parametric.Reg	0.5003	0.5924
12	OHE	PCA	Inverse.Reg	0.3842	0.6437
6	OHE	PLS	Inverse.Reg	0.3331	0.3999
15	OHE	PCA	Linear.Reg	0.3813	0.6428
7	**OHE**	**PLS**	**Linear.Reg**	**0.3258**	**0.4046**
9	OHE	PCA	Non.Parametric.Reg	0.4156	0.6474
3	**OHE**	**PLS**	**Non.Parametric.Reg**	**0.3700**	**0.3991**

Table 2: Performance Metrics in Total Expenditure Prediction with optimal d and $d = 1$. $N = 600$.

5.2 Performance in Classification: Line Poverty

In Tables 4 and 5 we show the different performance metrics for classification of poor households in the Greater Buenos Aires using a training sample of 400 and 250 households, respectively. In the first column we have the optimal dimension d^* for the reduction, obtained by maximizing the AUC value using 5-fold cross validation (CV). The second column shows the non-metric variables treatment method, the third column the method of reduction, and the fourth the predictive method to fit and classify. The fifth column shows the optimal 5-fold c^* that minimizes the Index of Union (IU) (4). The performance metrics (columns 6 to 9) are: the Misclassification Error (MCE), the True Positive Rate (TPR), the True Negative Rate (TNR), using c^*, and the average value of the 5-fold AUC for d^*. Finally, we add the optimal cut-off and performance metrics columns for $d = 1$. Analogous to the regression analysis, this allows evaluating which model methods provide the best results when building an interpretative SES index ($d = 1$) and inferring from real applications if there exists gains from increasing the dimension of reduction in the predictive performance (i.e. when $d^* > 1$). In addition, to evaluate the overall performance of the estimated models, we also consider the balance between the TPR and TNR. Although the expected levels of the exclusion or inclusion error depend on the policy goal, we assume that a balanced TPR-TNR is desirable and a measure of good

d^*	Treatment	Reduction	Model	$MSE(d^*)$	$MSE(d=1)$
15	MNMC	PCA	Inverse.Reg	0.4569	0.7307
15	MNMC	PLS	Inverse.Reg	0.4496	0.4768
9	MNMC	PCA	Linear.Reg	0.4656	0.7299
2	MNMC	PLS	Linear.Reg	0.4701	0.4764
2	MNMC	PCA	Non.Parametric.Reg	0.4948	0.7484
3	**MNMC**	**PLS**	**Non.Parametric.Reg**	**0.4674**	**0.4756**
12	NM	PCA	Inverse.Reg	0.7559	0.7575
1	NM	PLS	Inverse.Reg	0.9892	0.9892
1	NM	PCA	Linear.Reg	0.7563	0.7563
1	NM	PLS	Linear.Reg	0.9899	0.9899
1	**NM**	**PCA**	**Non.Parametric.Reg**	**0.7497**	**0.7497**
9	NM	PLS	Non.Parametric.Reg	0.8943	0.9941
1	NT cont	PCA	Inverse.Reg	0.5810	0.5810
1	**NT cont**	**PLS**	**Inverse.Reg**	**0.5703**	**0.5703**
1	NT cont	PCA	Linear.Reg	0.5804	0.5804
1	NT cont	PLS	Linear.Reg	0.5706	0.5706
2	NT cont	PCA	Non.Parametric.Reg	0.5973	0.6033
1	NT cont	PLS	Non.Parametric.Reg	0.5858	0.5858
2	NT mixed	PCA	Inverse.Reg	0.5970	0.6789
11	NT mixed	PLS	Inverse.Reg	0.6376	0.6854
2	NT mixed	PCA	Linear.Reg	0.5981	0.6784
10	NT mixed	PLS	Linear.Reg	0.6617	0.6782
4	**NT mixed**	**PCA**	**Non.Parametric.Reg**	**0.5600**	**0.6958**
15	NT mixed	PLS	Non.Parametric.Reg	0.6334	0.6904
11	OHE	PCA	Inverse.Reg	0.4338	0.7411
2	OHE	PLS	Inverse.Reg	0.4543	0.4737
15	OHE	PCA	Linear.Reg	0.4393	0.7437
3	**OHE**	**PLS**	**Linear.Reg**	**0.4467**	**0.4728**
9	OHE	PCA	Non.Parametric.Reg	0.5131	0.7424
2	OHE	PLS	Non.Parametric.Reg	0.4791	0.4792

Table 3: Performance Metrics in Total Expenditure Prediction with optimal d and $d = 1$. $N = 200$.

performance.

Assume that a typical interpretative index ($d = 1$) is desired. According to the 5-fold CV AUC metric, we can first observe that models whose reduction step is based on `PLS` achieve better results than ones based on `PCA`. Besides `PLS` as reduction method, the best models are those based on MNMC, NTcont, and OHE as treatment methods (from best to worst). However, while moderate results are obtained with NTmixed, with MNC the results are very poor. With respect to the other performance metrics, `PCA`-based models show relatively high MCE (in most cases greater than 0.5), and for every treatment method, `PLS`-based models yield lower MCE. In general, with `PLS` we obtain a more balanced TPR-TNR. However, when it is combined with NMC, the TPR and TNR are remarkably imbalanced (very low TPR), but this also happens with `PCA`-based indices. On one hand, if MNMC or OHE are used, the lowest inclusion errors (1-TNR) are obtained. On the other hand, if there is no treatment on the variables, we get the lowest exclusion error (1-TPR). Furthermore, in these cases it is also observed that there are no substantial differences between the prediction methods.
To summarize, for a typical SES-index, `PLS` as a reduction method has higher predictive power than `PCA`. MNMC yields a high AUC, low MCE, and a more balanced TPR-TNR. More precisely, it tends to maximize the TNR. This means that a targeted program using it would achieve greater cost efficiency.

If we increase the number of components in the reduction step, there are evident gains in terms of predictive power. For the optimal d^*, results for `PCA` and `PLS`-based models are now comparable. However, as in the case of the regression models, `PCA` requires a higher number of components than `PLS` to achieve its maximum AUC or minimum MCE. With respect to the treatment methods, despite the improvement, NMC still exhibits relatively poor results. Now, OHE outperforms NTcont in terms of AUC and MCE, which was not the case when $d = 1$. NTcont competes with MNMC but, on average, MNMC gives a simpler index in terms of lower d^*. NTmixed produces relatively modest results. For every model, both TPR and TNR improve and their balance is more noticeable. Nonetheless, contrary to the case with $d = 1$, it is not possible to determine which models minimise the inclusion error or the exclusion error. In other words, no clear patters are revealed in this sense. For the same treatment-reduction methods, which metric (TPR or TNR) is higher depends on the prediction method (Logistic, Inverse regression,

LDA, or QDA).
In broad terms, searching for the optimal d^* does improve the predictive power of models, at the cost of sacrificing interpretability. Compared to the case with $d = 1$, `PCA`based models benefit most from increasing the number of components, similarly to the regression example. Finally, the ultimate selection of the model will depend on which measure is to be maximized (TPR or TNR).
Table 5 exhibits the same performance metrics for household poverty classification as in Table 4, but with a smaller sample ($N = 250$). These results show the same patterns but with lower predictive power than the $N = 400$ case. However, in this case the TPR and TNR are more balanced (although with smaller valeus than in with $N = 400$).

Therefore, if the goal is to have an index that not only has the best predictive power but also has the smallest possible dimension, given the advantages of a more synthetic index (e.g. interpretation of projections and weights), `PLS` with MNMC or OHE treatments is a good choice. With these treatment-reduction methodologies, it is possible to have two-dimensional SES indices with the best classification results.

				with d^*					with $d = 1$				
d^*	Treatment	Reduction	Prediction Method	cut-off	MCE	TPR	TNR	AUC	cut-off	MCE	TPR	TNR	AUC
12	MNMC	PCA	Inverse.Reg	0.1880	0.1803	0.7409	0.8405	0.8657	0.2060	0.5149	0.4596	0.4849	0.5090
2	MNMC	PLS	Inverse.Reg	0.3280	0.2107	0.7995	0.7858	0.8530	0.3060	0.2379	0.7159	0.7726	0.8072
12	**MNMC**	**PCA**	**LDA**	**0.1880**	**0.1803**	**0.7409**	**0.8405**	**0.8657**	**0.2060**	**0.5149**	**0.4596**	**0.4849**	**0.5090**
2	MNMC	PLS	LDA	0.3020	0.2132	0.7995	0.7829	0.8530	0.3060	0.2379	0.7159	0.7726	0.8072
6	MNMC	PCA	Logistic.Reg	0.2820	0.2030	0.8080	0.7912	0.8681	0.2060	0.5149	0.4596	0.4849	0.5090
2	**MNMC**	**PLS**	**Logistic.Reg**	**0.3300**	**0.2157**	**0.8100**	**0.7758**	**0.8532**	**0.3120**	**0.2276**	**0.7034**	**0.7893**	**0.8072**
4	MNMC	PCA	QDA	0.2640	0.2293	0.8080	0.7582	0.8537	0.2040	0.5436	0.4923	0.4444	0.5214
2	MNMC	PLS	QDA	0.2820	0.2132	0.7995	0.7829	0.8547	0.2940	0.2304	0.7159	0.7824	0.8072
7	**NM**	**PCA**	**Inverse.Reg**	**0.1040**	**0.1764**	**0.3300**	**0.9512**	**0.7925**	**0.1160**	**0.7911**	**0.9875**	**0.0033**	**0.4036**
16	NM	PLS	Inverse.Reg	0.1000	0.3210	0.0250	0.7108	0.5387	0.1000	0.2089	0.0000	1.0000	0.3567
7	NM	PCA	LDA	0.1020	0.1764	0.3405	0.9479	0.7925	0.1160	0.7911	0.9875	0.0033	0.4036
3	NM	PLS	LDA	0.1000	0.2989	0.1258	0.8529	0.4484	0.1000	0.2089	0.0000	1.0000	0.3567
8	NM	PCA	Logistic.Reg	0.1060	0.1909	0.3453	0.9296	0.7707	0.1160	0.7937	0.9625	0.0066	0.4036
17	NM	PLS	Logistic.Reg	0.1000	0.2912	0.0980	0.8725	0.4729	0.1000	0.2089	0.0000	1.0000	0.3567
8	NM	PCA	QDA	0.2820	0.3560	0.5093	0.6743	0.6316	0.1000	0.7911	1.0000	0.0000	0.4859
6	NM	PLS	QDA	0.7860	0.3942	0.7245	0.5748	0.7030	0.1000	0.2089	0.0000	1.0000	0.3567
3	NT cont	PCA	Inverse.Reg	0.1740	0.1888	0.8282	0.8065	0.8658	0.2080	0.5268	0.4740	0.4694	0.4754
20	NT cont	PLS	Inverse.Reg	0.1400	0.1949	0.7315	0.8271	0.8476	0.2040	0.2615	0.7881	0.7275	0.8038
3	NT cont	PCA	LDA	0.1740	0.1888	0.8282	0.8065	0.8658	0.2080	0.5268	0.4740	0.4694	0.4754
20	NT cont	PLS	LDA	0.1400	0.1949	0.7315	0.8271	0.8476	0.2260	0.2522	0.7506	0.7472	0.8038
3	**NT cont**	**PCA**	**Logistic.Reg**	**0.1900**	**0.1969**	**0.8282**	**0.7955**	**0.8635**	**0.2080**	**0.5268**	**0.4740**	**0.4694**	**0.4754**
20	**NT cont**	**PLS**	**Logistic.Reg**	**0.1920**	**0.2116**	**0.8080**	**0.7834**	**0.8493**	**0.2060**	**0.2615**	**0.7881**	**0.7275**	**0.8038**
1	NT cont	PCA	QDA	0.2080	0.5389	0.4260	0.4677	0.4563	0.2080	0.5389	0.4260	0.4677	0.4563
19	NT cont	PLS	QDA	0.1560	0.1845	0.6566	0.8600	0.8531	0.2160	0.2547	0.7506	0.7442	0.8038
8	**NT mixed**	**PCA**	**Inverse.Reg**	**0.2000**	**0.1747**	**0.7966**	**0.8346**	**0.8699**	**0.2140**	**0.3884**	**0.6795**	**0.5913**	**0.6997**
1	**NT mixed**	**PLS**	**Inverse.Reg**	**0.2080**	**0.4652**	**0.5510**	**0.5332**	**0.6057**	**0.2080**	**0.4652**	**0.5510**	**0.5332**	**0.6057**
8	NT mixed	PCA	LDA	0.2000	0.1747	0.7966	0.8346	0.8699	0.2140	0.3884	0.6795	0.5913	0.6997
1	NT mixed	PLS	LDA	0.2080	0.4652	0.5510	0.5332	0.6057	0.2080	0.4652	0.5510	0.5332	0.6057
8	NT mixed	PCA	Logistic.Reg	0.2140	0.1916	0.7946	0.8121	0.8655	0.2120	0.3859	0.6795	0.5946	0.6997
1	NT mixed	PLS	Logistic.Reg	0.2080	0.4652	0.5510	0.5332	0.6057	0.2080	0.4652	0.5510	0.5332	0.6057
5	NT mixed	PCA	QDA	0.1960	0.1949	0.7736	0.8145	0.8592	0.2180	0.3834	0.6900	0.5943	0.6997
1	NT mixed	PLS	QDA	0.1900	0.4484	0.5385	0.5568	0.5917	0.1900	0.4484	0.5385	0.5568	0.5917
17	OHE	PCA	Inverse.Reg	0.1700	0.1617	0.7909	0.8527	0.8746	0.2040	0.5349	0.4951	0.4514	0.4999
2	OHE	PLS	Inverse.Reg	0.1760	0.1837	0.8387	0.8093	0.8774	0.2160	0.2642	0.7085	0.7470	0.7855
17	OHE	PCA	LDA	0.1700	0.1617	0.7909	0.8527	0.8746	0.2040	0.5349	0.4951	0.4514	0.4999
2	OHE	PLS	LDA	0.1760	0.1837	0.8387	0.8093	0.8774	0.2160	0.2667	0.7085	0.7441	0.7855
17	**OHE**	**PCA**	**Logistic.Reg**	**0.1960**	**0.1812**	**0.8541**	**0.8075**	**0.8785**	**0.2040**	**0.5349**	**0.4951**	**0.4514**	**0.4999**
2	**OHE**	**PLS**	**Logistic.Reg**	**0.2060**	**0.1836**	**0.8387**	**0.8097**	**0.8782**	**0.2260**	**0.2667**	**0.7085**	**0.7441**	**0.7855**
2	**OHE**	**PCA**	**QDA**	**0.1800**	**0.1969**	**0.7696**	**0.8126**	**0.8620**	**0.2060**	**0.5518**	**0.4923**	**0.4362**	**0.4926**
2	**OHE**	**PLS**	**QDA**	**0.1680**	**0.1890**	**0.8282**	**0.8057**	**0.8740**	**0.2040**	**0.2767**	**0.7295**	**0.7241**	**0.7855**

Table 4: Performance Metrics in Monetary Poverty Classification with optimal d and $N = 400$.

				with d^*					with $d = 1$				
d	Treatment	Reduction	Prediction Method	c	MCE	TPR	TNR	AUC	c	MCE	TPR	TNR	AUC
12	MNMC	PCA	Inverse.Reg	0.4100	0.2395	0.7214	0.7617	0.7833	0.2060	0.5664	0.5246	0.4036	0.4857
2	MNMC	PLS	Inverse.Reg	0.5100	0.2392	0.7701	0.7475	0.7928	0.4540	0.2555	0.7427	0.7366	0.7684
12	MNMC	PCA	LDA	0.4100	0.2395	0.7214	0.7617	0.7833	0.2060	0.5664	0.5246	0.4036	0.4857
2	MNMC	PLS	LDA	0.5100	0.2392	0.7701	0.7475	0.7928	0.4520	0.2593	0.7427	0.7318	0.7684
6	MNMC	PCA	Logistic.Reg	0.5280	0.2342	0.7701	0.7560	0.8035	0.2060	0.5664	0.5246	0.4036	0.4857
2	MNMC	PLS	Logistic.Reg	0.5140	0.2392	0.7547	0.7523	0.7921	0.4400	0.2593	0.7427	0.7318	0.7684
4	**MNMC**	**PCA**	**QDA**	**0.4480**	**0.2393**	**0.7901**	**0.7425**	**0.8034**	**0.2060**	**0.5415**	**0.3919**	**0.4686**	**0.4145**
2	**MNMC**	**PLS**	**QDA**	**0.5160**	**0.2315**	**0.7547**	**0.7619**	**0.7952**	**0.4400**	**0.2593**	**0.7427**	**0.7318**	**0.7684**
7	**NM**	**PCA**	**Inverse.Reg**	**0.1000**	**0.1817**	**0.2024**	**0.9836**	**0.6668**	**0.1000**	**0.7897**	**1.0000**	**0.0000**	**0.4095**
16	NM	PLS	Inverse.Reg	0.1400	0.3538	0.1062	0.6218	0.4953	0.1120	0.3280	0.0600	0.8390	0.3228
7	NM	PCA	LDA	0.1000	0.1817	0.2024	0.9836	0.6668	0.1000	0.7897	1.0000	0.0000	0.4095
3	NM	PLS	LDA	0.1000	0.3441	0.1105	0.8087	0.3989	0.1120	0.3280	0.0600	0.8390	0.3228
8	NM	PCA	Logistic.Reg	0.1000	0.1876	0.2357	0.9670	0.7120	0.1000	0.7897	1.0000	0.0000	0.4095
17	NM	PLS	Logistic.Reg	0.1000	0.3245	0.1347	0.8212	0.3311	0.1140	0.3201	0.0600	0.8488	0.3228
8	NM	PCA	QDA	0.1220	0.2917	0.4579	0.7768	0.7106	0.1000	0.7897	1.0000	0.0000	0.5422
6	NM	PLS	QDA	0.3700	0.4810	0.5513	0.5141	0.5609	0.1100	0.3084	0.0600	0.8634	0.3276
3	NT cont	PCA	Inverse.Reg	0.1940	0.2036	0.8341	0.7858	0.8518	0.2100	0.6392	0.5148	0.3223	0.4496
20	NT cont	PLS	Inverse.Reg	0.1060	0.2413	0.7448	0.7628	0.8093	0.2340	0.2846	0.7438	0.7091	0.7811
3	NT cont	PCA	LDA	0.1940	0.2036	0.8341	0.7858	0.8518	0.2100	0.6392	0.5148	0.3223	0.4496
20	NT cont	PLS	LDA	0.1060	0.2413	0.7448	0.7628	0.8093	0.2340	0.2846	0.7438	0.7091	0.7811
3	NT cont	PCA	Logistic.Reg	0.2140	0.2075	0.8141	0.7858	0.8496	0.2100	0.6432	0.5148	0.3170	0.4496
20	NT cont	PLS	Logistic.Reg	0.2240	0.2310	0.7238	0.7820	0.8164	0.2420	0.2767	0.7238	0.7238	0.7811
1	**NT cont**	**PCA**	**QDA**	**0.2100**	**0.5865**	**0.4505**	**0.4130**	**0.3885**	**0.2100**	**0.5865**	**0.4505**	**0.4130**	**0.3885**
19	**NT cont**	**PLS**	**QDA**	**0.1000**	**0.2381**	**0.5076**	**0.8327**	**0.7614**	**0.2380**	**0.2728**	**0.7238**	**0.7287**	**0.7811**
8	**NT mixed**	**PCA**	**Inverse.Reg**	**0.1920**	**0.1939**	**0.7762**	**0.8149**	**0.8560**	**0.2200**	**0.3850**	**0.6545**	**0.6047**	**0.6917**
1	**NT mixed**	**PLS**	**Inverse.Reg**	**0.2080**	**0.3970**	**0.6211**	**0.5965**	**0.6266**	**0.2080**	**0.3970**	**0.6211**	**0.5965**	**0.6266**
8	NT mixed	PCA	LDA	0.1920	0.1939	0.7762	0.8149	0.8560	0.2180	0.3930	0.6712	0.5890	0.6917
1	NT mixed	PLS	LDA	0.2080	0.3970	0.6211	0.5965	0.6266	0.2080	0.3970	0.6211	0.5965	0.6266
8	NT mixed	PCA	Logistic.Reg	0.1940	0.2123	0.7901	0.7851	0.8504	0.2200	0.3892	0.6712	0.5937	0.6917
1	NT mixed	PLS	Logistic.Reg	0.2120	0.3857	0.6057	0.6165	0.6266	0.2120	0.3857	0.6057	0.6165	0.6266
5	NT mixed	PCA	QDA	0.2280	0.2015	0.7987	0.7988	0.8454	0.2300	0.3890	0.6545	0.5995	0.6902
1	NT mixed	PLS	QDA	0.1780	0.4003	0.6143	0.5965	0.6402	0.1780	0.4003	0.6143	0.5965	0.6402
17	OHE	PCA	Inverse.Reg	0.1760	0.2232	0.7300	0.7924	0.8279	0.1840	0.6259	0.5974	0.3140	0.4971
2	OHE	PLS	Inverse.Reg	0.1780	0.2032	0.7959	0.7954	0.8549	0.2060	0.2618	0.7112	0.7459	0.7763
17	OHE	PCA	LDA	0.1760	0.2232	0.7300	0.7924	0.8279	0.1840	0.6259	0.5974	0.3140	0.4971
2	OHE	PLS	LDA	0.1760	0.2032	0.7959	0.7954	0.8549	0.2060	0.2618	0.7112	0.7459	0.7763
17	**OHE**	**PCA**	**Logistic.Reg**	**0.1520**	**0.2299**	**0.7701**	**0.7689**	**0.8337**	**0.1840**	**0.6259**	**0.5974**	**0.3140**	**0.4971**
2	**OHE**	**PLS**	**Logistic.Reg**	**0.1960**	**0.2145**	**0.8141**	**0.7764**	**0.8535**	**0.2180**	**0.2657**	**0.7112**	**0.7410**	**0.7763**
2	**OHE**	**PCA**	**QDA**	**0.1620**	**0.2275**	**0.8294**	**0.7577**	**0.8499**	**0.2100**	**0.6105**	**0.3934**	**0.3926**	**0.3875**
2	**OHE**	**PLS**	**QDA**	**0.1720**	**0.2069**	**0.7959**	**0.7904**	**0.8453**	**0.1960**	**0.2657**	**0.7112**	**0.7410**	**0.7763**

Table 5: Performance Metrics in Monetary Poverty Classification with optimal d and $N = 250$.

6 Conclusion

In this paper, we compared the predictive power of reduced-dimension SES indices over a set of metric and non-metric variables, using different treatment procedures for the non-metric variables and alternative prediction methods. This was performed for both regression and classification problems. For treating non-metric variables, we applied techniques commonly used in the related literature — NMC, OHE, NTcont, NTmixed —, proposing a multivariate extension of the normal mean coding (MNMC) as a new method that captures the multivariate nature of the mixed predictor variables used in the SES index construction. The prediction methods used were Linear, Non-linear, and Inverse Regression for regression problems, and Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Logistic Regression for classification ones.

To evaluate the reduction/treatment/prediction methods, first we worked with five simulation settings, that take into account different issues as linear and non-linear relationships between variables (response and variables), the dimension of the reduction and different sample sizes.

Results show that MNMC was the best method for treating non-metric variables both in regression and classification simulations. In regression problems, the results showed that the inclusion of dummies through the OHE is a good strategy too. when non-linearity is present, the performance of the linear model drops and non-parametric prediction methods showed the lowest MSE in regression problems. Although the results by reduction methods had barely any differences, `PCA` displayed lower MSE. In the linear settings not much efficiency is lost if inverse regression or non-parametric prediction is used. As we expected, if we use only one of the two true projections, `PLS` gives better results than `PCA`. In classification problems, logistic regression and LDA presented similar results. Due to the fact that the simulated models have constant variables, QDA applies but since it requires the estimation of more parameters the results are not as good as when we apply LDA.

Next, we worked with two real data sets: the Argentinian National Household Expenditure Survey of 2018 (Encuesta Nacional de Gastos de los Hogares-ENGHO), and the Argentinian Household Survey (Encuesta Permanente de Hogares-EPH) for the second quarter of 2017, both for Greater Buenos Aires (GBA) demographic area, in two sample sizes. In this application we also studied the gains of increasing the number of components ($d > 1$) in the performance of a composite, if interpretability of an SES index is not sought.

The regression application was done over ENGHO, and had as a response variable the logarithm of the household total expenditure. Here when $d = 1$, the best models are those based on MNMC and OHE as treatment and `PLS` as reduction method, with negligible differences by prediction method. We found the optimal dimension of reduction minimizing the MSE (d^*) running a 5-fold CV. With $d > 1$, `PCA` reached similar as `PLS`But `PCA` needs more components than `PLS` to achieve the lowest MSE, so we showed that the latter is more parsimonious in terms of constructing a composite index. In the smaller data sample, the overall performances were poorer but presented patterns similar to the bigger data sample.

For the EPH data set, a classifier was trained using poor and non-poor households as a response variables. Similar to the regression problem, we used a 5-fold CV to find the optimal dimension of reduction that maximizes the AUC value (d^*). With $d = 1$, the best models are those based on MNMC, NTcont, and OHE as treatment methods (in descendent order) and `PLS` as a reduction method, in terms of higher AUC, lower MCE and better TPR-TNR balance. `PLS`-MNMC maximized the TNR, which means using this combination in a targeted program would achieve greater cost efficiency. As in the regression problem, with $d > 1$ `PCA` reached similar results as `PLS`, but requiring more components. Over the treatments, MNMC was more parsimonious in terms of lower d^*, improving the TPR-TNR balance. Nevertheless, for the same treatment-reduction methods, which metric (TPR or TNR) is higher depends now on the prediction method. Again, with a smaller sample size we could see the same patterns with lower predictive power, but with a better TPR-TNR balance.

In broad terms, we can highlight two main findings. First of all, we show that for predictive purposes, there is a clear gain in increasing the number of factors. Second, when selecting an optimal d^* with 5-fold CV, `PLS` is more parsimonious than `PCA` regarding the number of projections to maximize the AUC.

A Box-Plots of Performance Metrics Results from Simulations

A.1 Performance in Regression

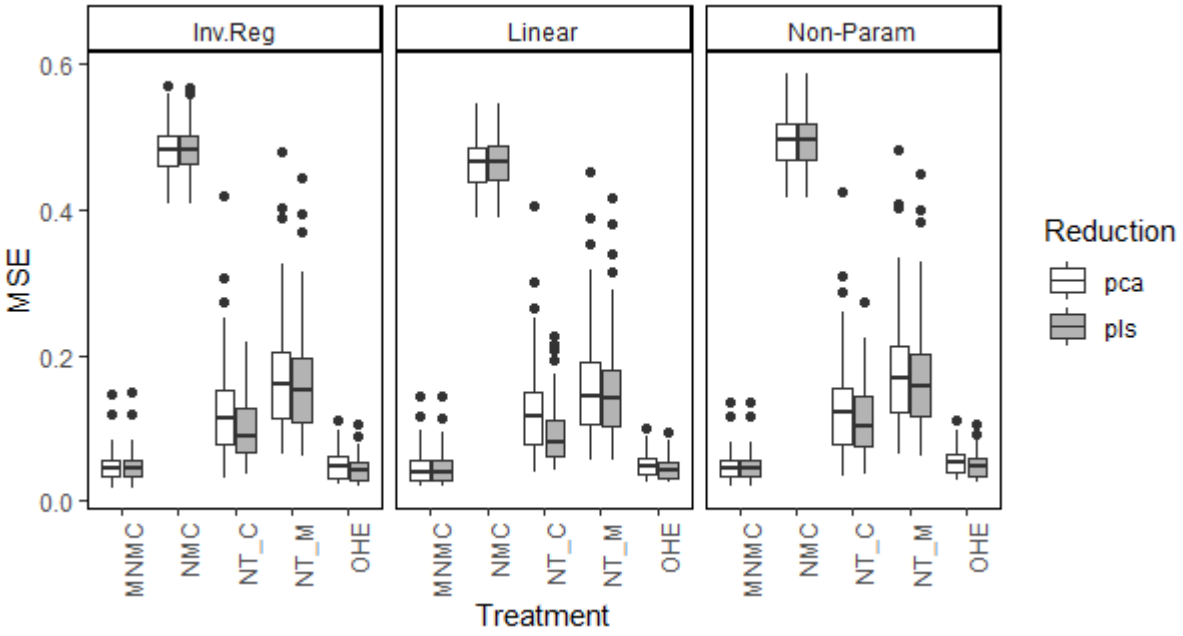


Figure 3: MSE in regression for setting **S1** with $n = 100$ using true d

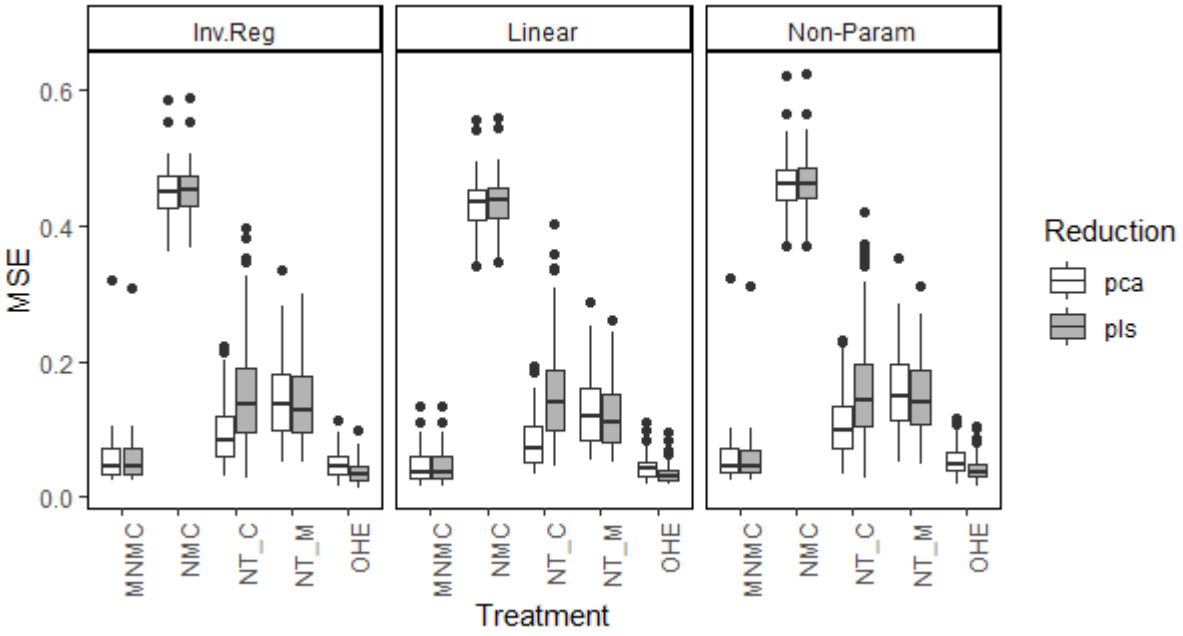


Figure 4: MSE in regression for setting **S2** with $n = 100$ using true d

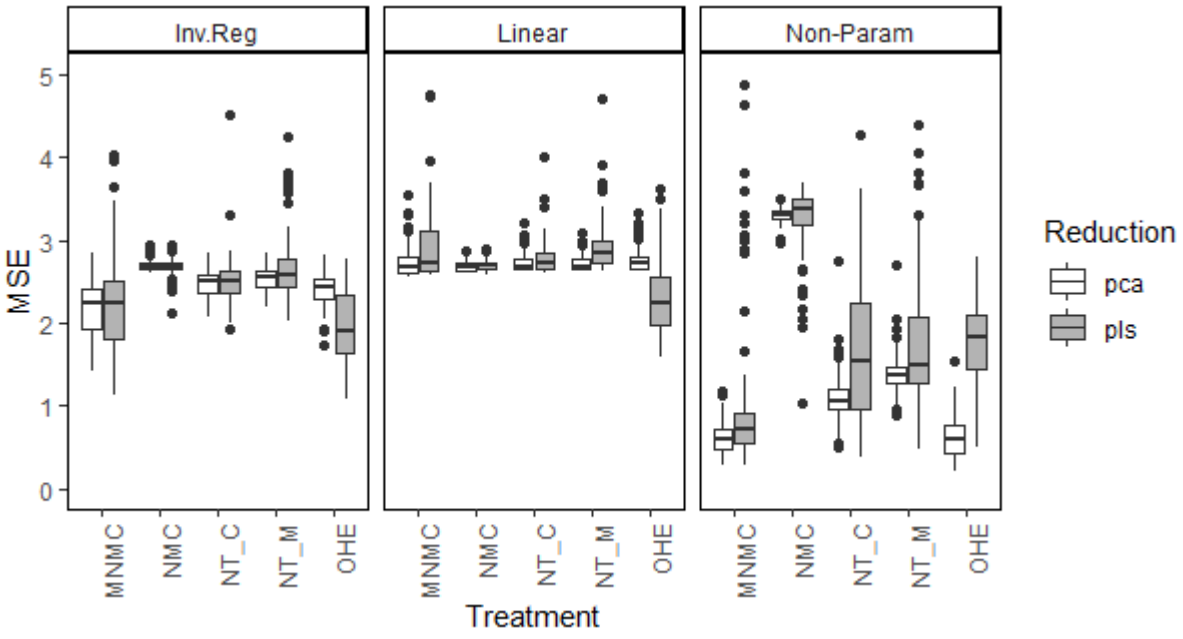


Figure 5: MSE in regression for setting **S3** with $n = 100$ using true d

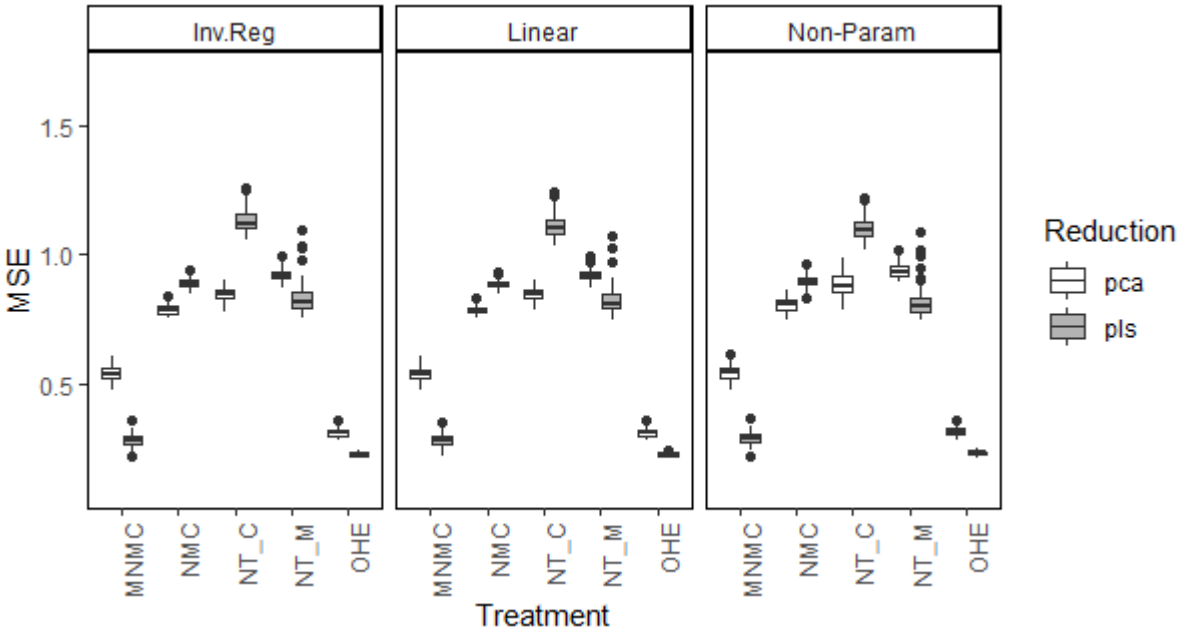


Figure 6: MSE in regression for setting **S4** with $n = 1000$ using $d = 1$ when true $d = 2$

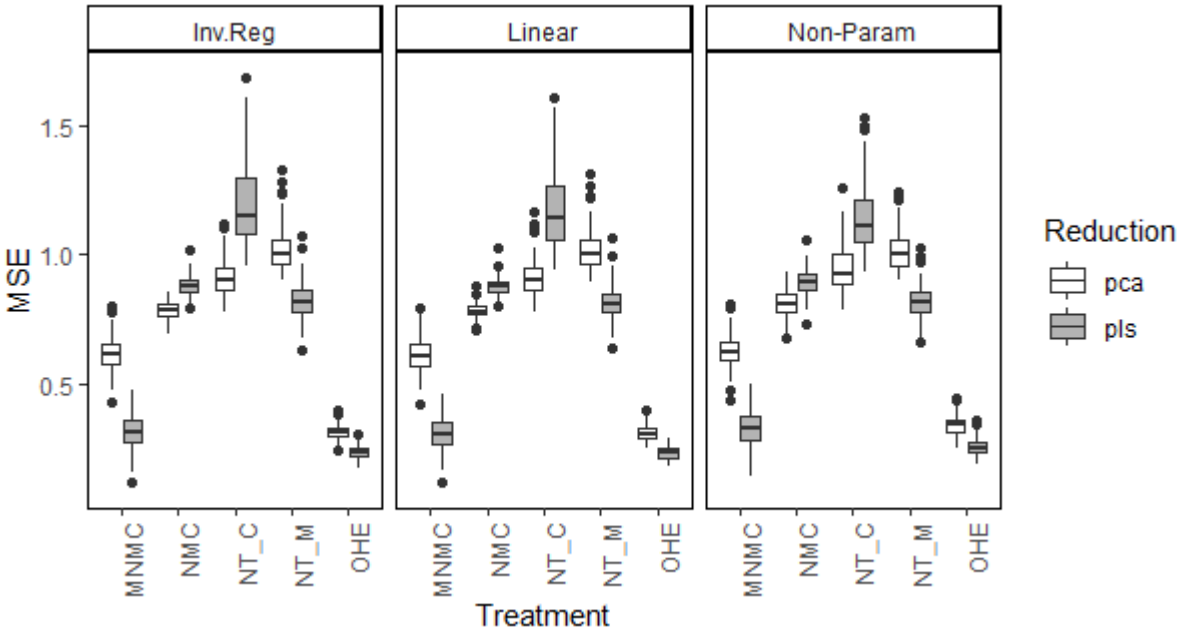


Figure 7: MSE in regression for setting **S4** with $n = 100$ using $d = 1$ when true $d = 2$

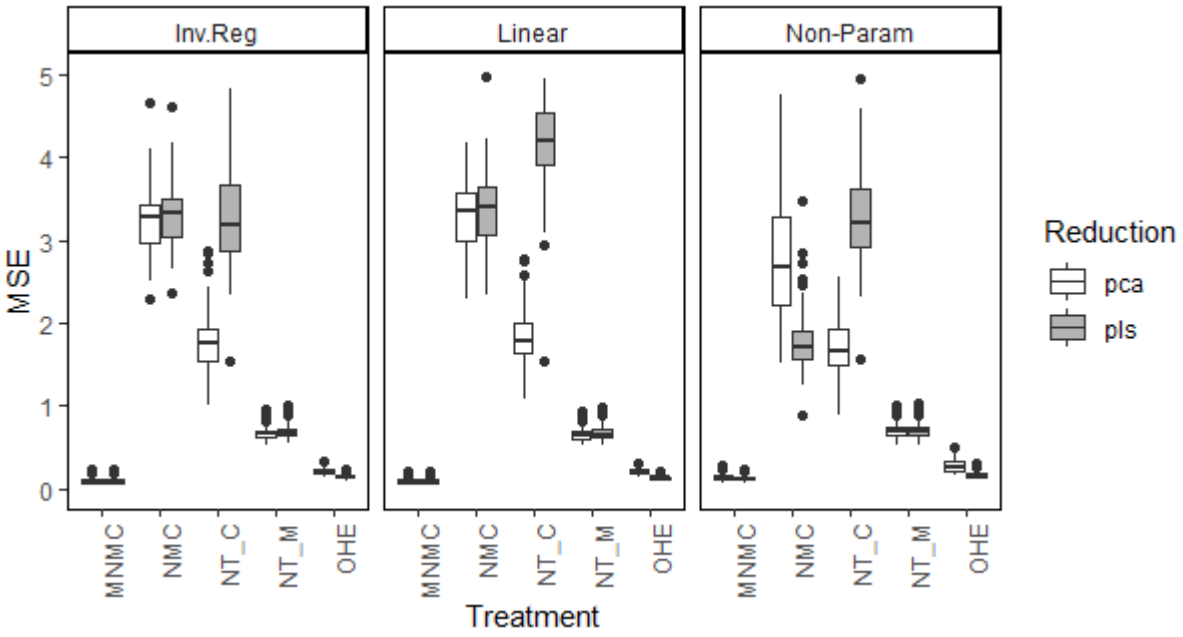


Figure 8: MSE in regression for setting **S4** with $n = 100$ using true $d = 2$

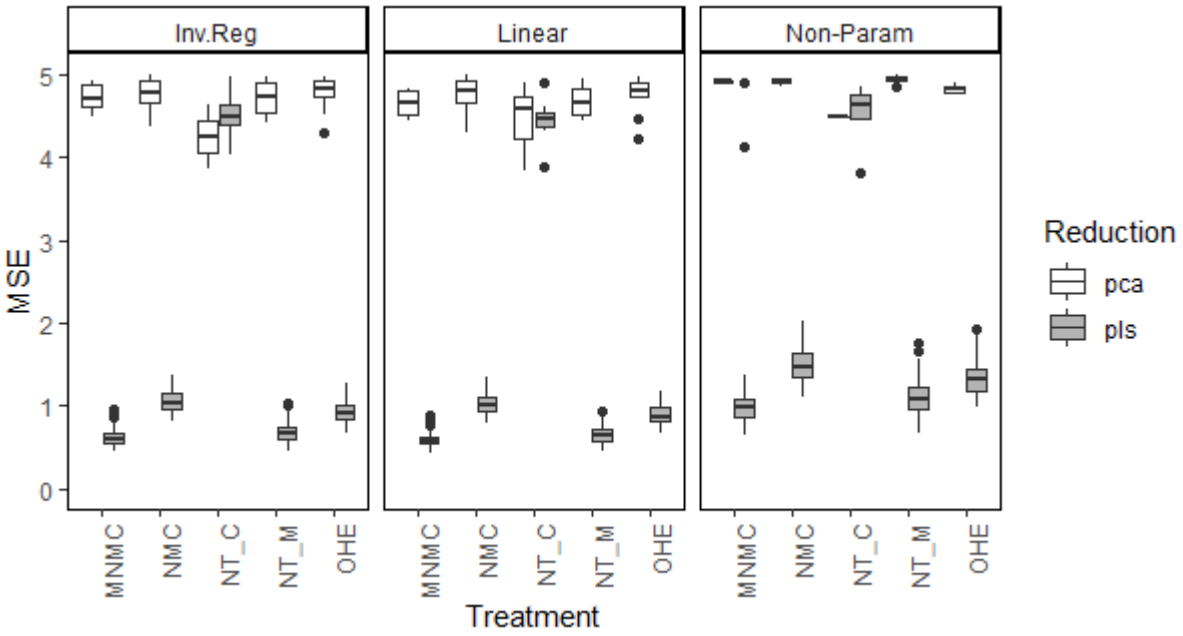


Figure 9: MSE in regression for setting **S5** with $n = 100$ using $d = 3$

A.2 Performance in Classification

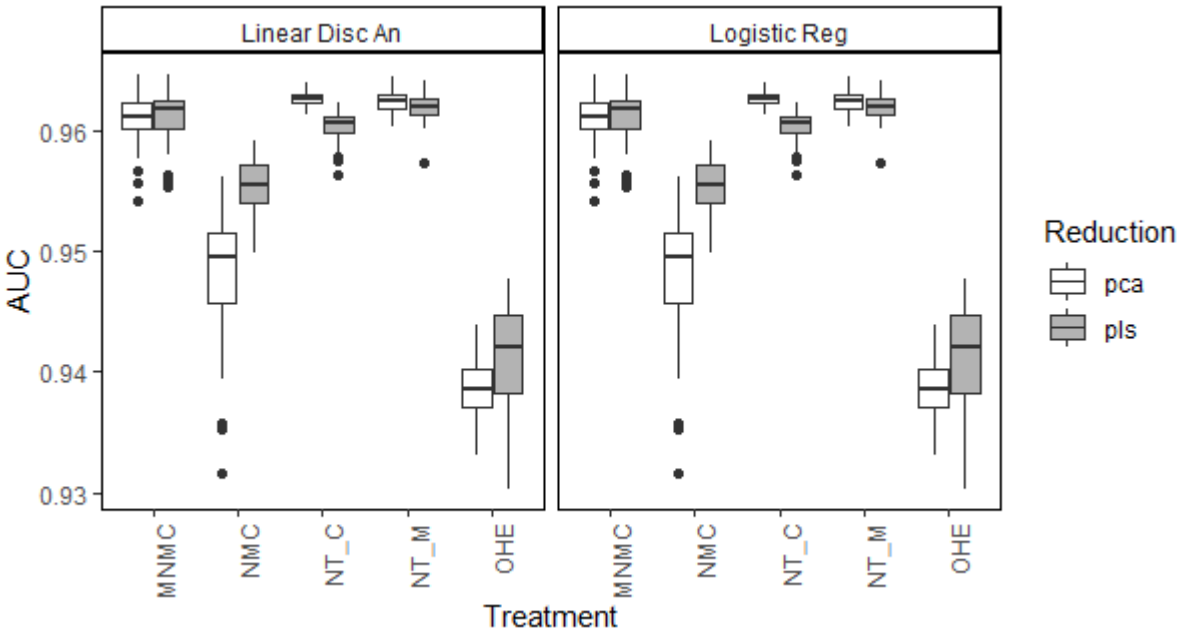


Figure 10: AUC in classification for setting **S1** with $n = 100$ using true d

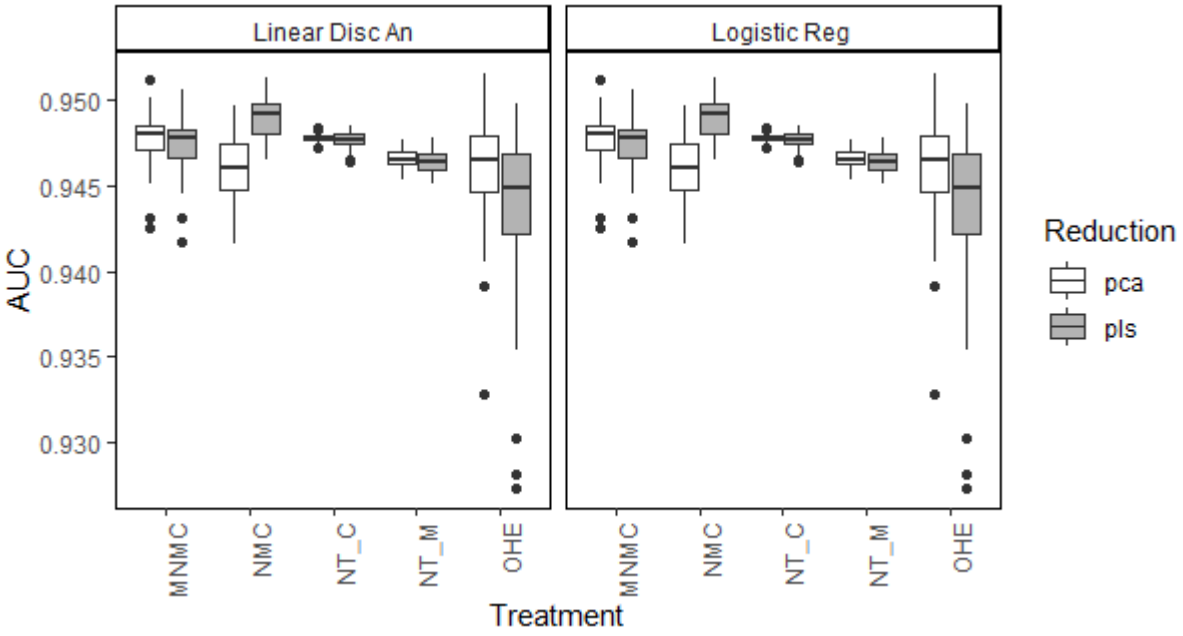


Figure 11: AUC in classification for setting **S2** with $n = 100$ using true d

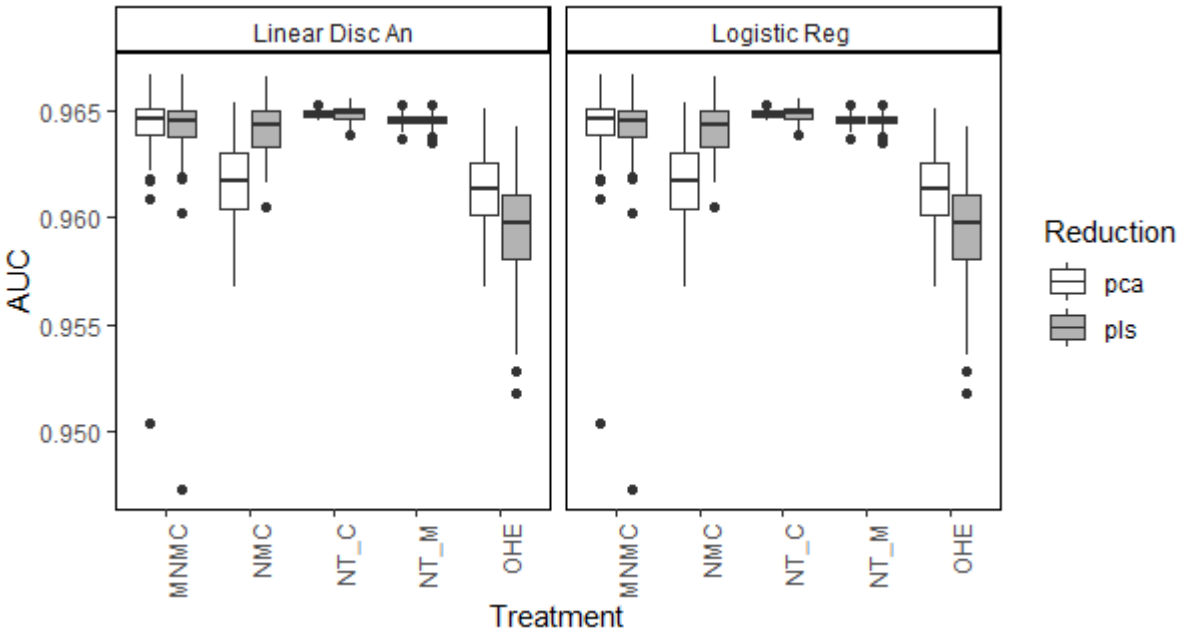


Figure 12: AUC in classification for setting **S3** with $n = 100$ using true d

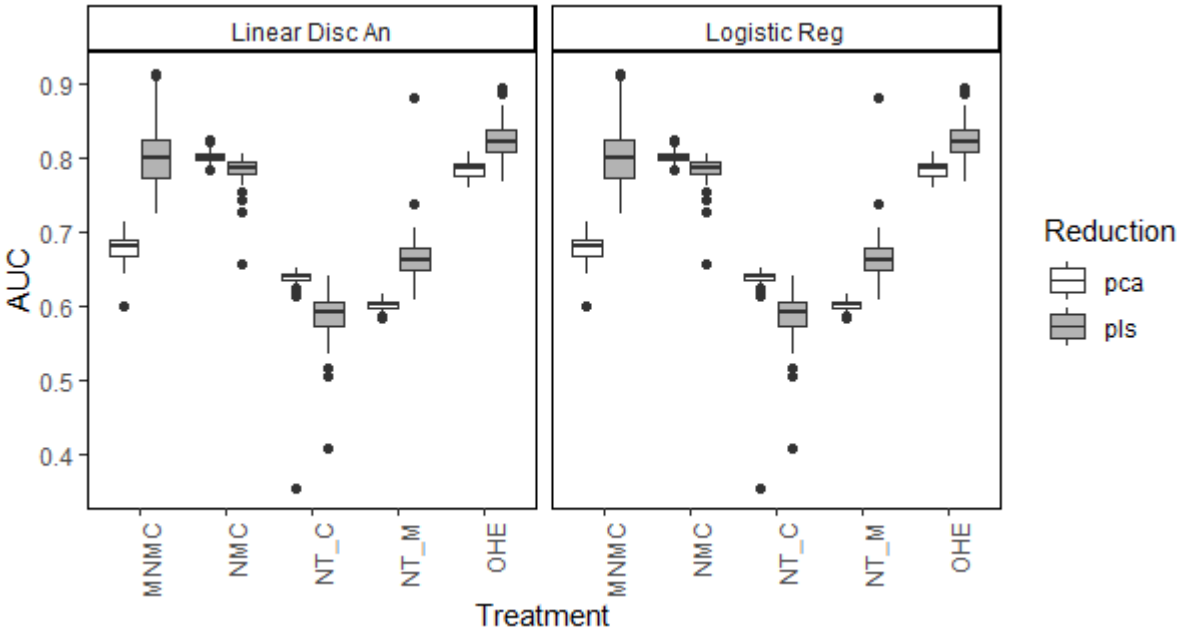


Figure 13: AUC in classification for setting **S4** with $n = 100$ using $d = 1$ when true $d = 2$

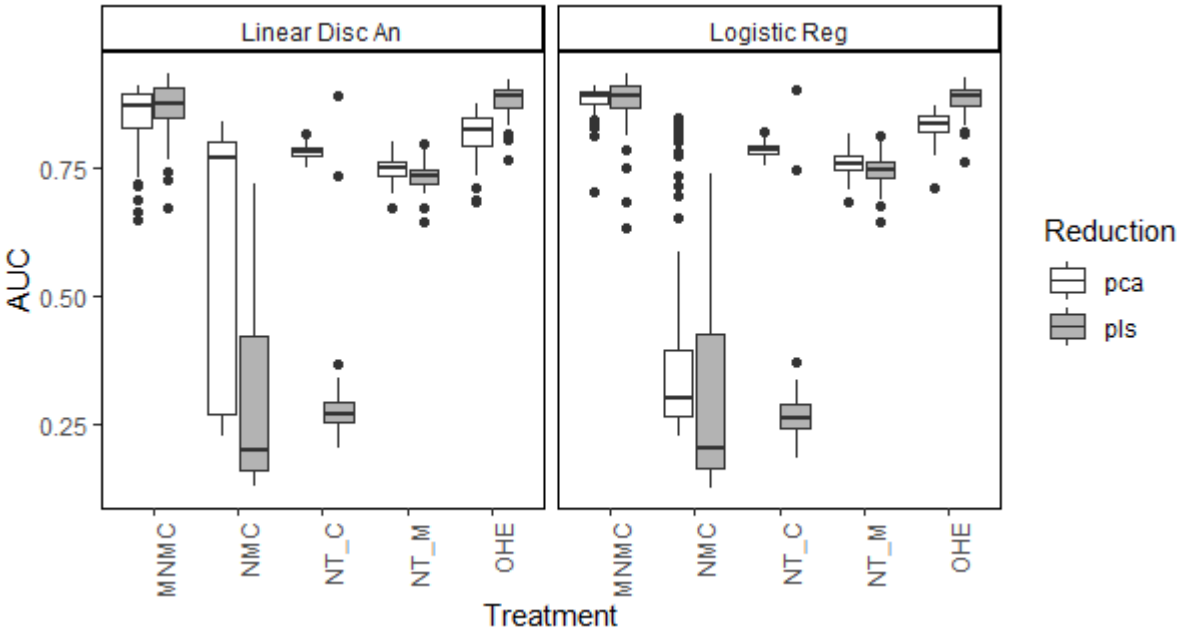


Figure 14: AUC in classification for setting **S4** with $n = 100$ using true $d = 2$

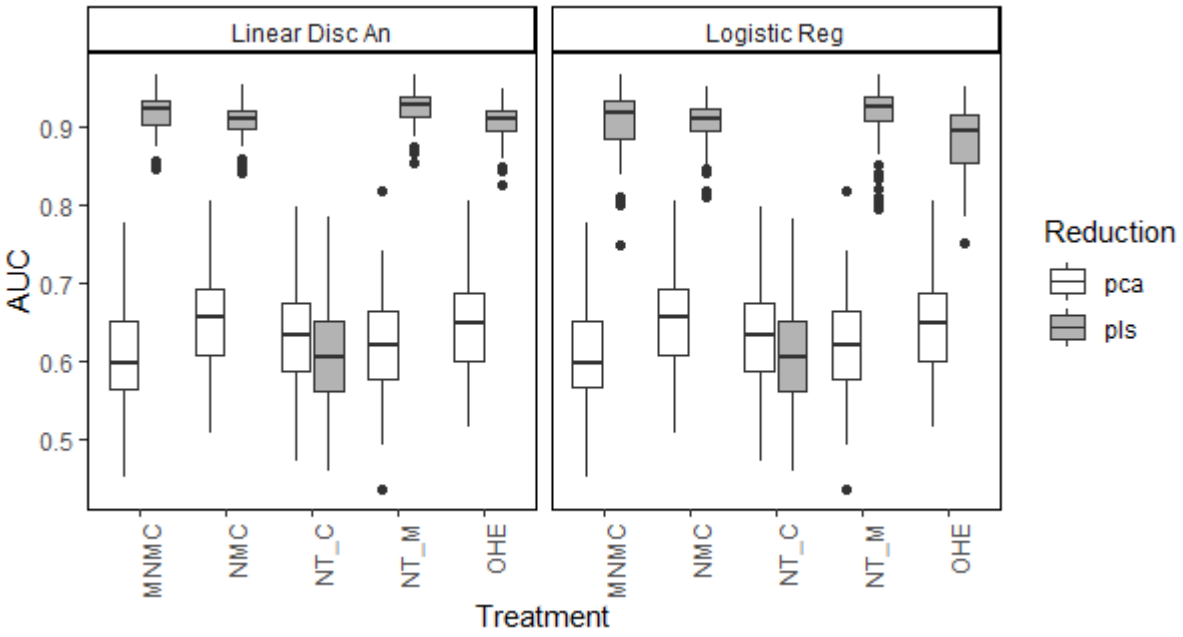


Figure 15: AUC in classification for setting **S5** with $n = 100$ using $d = 3$

References

[1] Adragni, K. and Cook, R. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

[2] Chin, W. W., Marcolin, B. L., and Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a monte carlo simulation study and an electronic-mail emotion/adoption study. *Information systems research*, 14(2):189–217.

[3] Coady, D., Grosh, M. E., and Hoddinott, J. (2004). Targeting of transfers in developing countries: Review of lessons and experience.

[4] Cook, R. D. and Forzani, L. (2018). Big data and partial least-squares prediction. *Canadian Journal of Statistics*, 46(1):62–78.

[5] Cook, R. D. and Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics*, 47(2):884–908.

[6] Cook, R. D. and Forzani, L. (2020). Envelopes: A new chapter in partial least squares regression. *Journal of Chemometrics*, 34(10):e3287.

[7] Cook, R. D. and Forzani, L. (2021). Pls regression algorithms in the presence of nonlinearity. *Chemometrics and Intelligent Laboratory Systems*, 213:104307.

[8] Cowan, C. D., Hauser, R. M., Kominski, R. A., Levin, H. M., Lucas, S. R., Morgan, S. L., and Chapman, C. (2012). Improving the measurement of socioeconomic status for the national assessment of educational progress: A theoretical foundation. *National Center for Education Statistics*, 2012.

[9] Deer, L. K., Shields, G. S., Alen, N. V., and Hostinar, C. E. (2021). Curvilinear associations between family income in early childhood and the cortisol awakening response in adolescence. *Psychoneuroendocrinology*, 129:105237.

[10] Duarte, S., Forzani, L., García Arancibia, R., Llop, P., and Tomassi, D. (2021a). Socioeconomic index for income and poverty prediction: A sufficient dimension reduction approach. *Review of Income and Wealth*.

[11] Duarte, S., Forzani, L., García Arancibia, R., Llop, P., and Tomassi, D. (2021b). Socioeconomic index for income and poverty prediction: A sufficient dimension reduction approach. *Review of Income and Wealth*.

[12] Earnest, A., Ong, M. E., Shahidah, N., Chan, A., Wah, W., and Thumboo, J. (2015). Derivation of indices of socioeconomic status for health services research in asia. *Preventive Medicine Reports*, 2:326–332.

[13] Filmer, D. and Pritchett, J. (2001). Estimating wealth effect without expenditure data -of tears: An application to educational enrollments in states of india. *Demography*, 38(4):115–132.

[14] Forzani, L., Arancibia, R. G., Llop, P., and Tomassi, D. (2018a). Supervised dimension reduction for ordinal predictors. *Computational Statistics & Data Analysis*, 125:136–155.

[15] Forzani, L., García-Arancibia, R., Llop, P., and Tomassi, D. (2018b). Supervised dimension reduction for ordinal predictors. *Computational Statistics and Data Analysis*, 125.

[Forzani et al.] Forzani, L., Rodriguez, D., and Sued, M.

[17] Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models*. CRC Press, Boca Raton, FL.

[18] Gwatkin, D. R., Rutstein, S., Johnson, K., Suliman, E., Wagstaff, A., and Amouzou, A. (2007). Socio-economic differences in health, nutrition, and population. *Washington, DC: The World Bank*, pages 1–301.

[19] Hancock, J. T. and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):1–41.

[20] Hanna, R. and Olken, B. A. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–26.

[21] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[22] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[23] Houssou, N., Zeller, M., Alcaraz V., G., Schwarze, S., and Johannsen, J. (2007). Proxy means tests for targeting the poorest households – applications to uganda. 106th Seminar, October 25-27, 2007, Montpellier, France 7946, European Association of Agricultural Economists.

[24] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[25] Jolliffe, I. (2002). *Principal Component Analysis, Second Edition*. Springer, New York.

[26] Khodayari Moez, E., Maximova, K., Sim, S., Senthilselvan, A., and Pabayo, R. (2022). Developing a socioeconomic status index for chronic disease prevention research in canada. *International Journal of Environmental Research and Public Health*, 19(13).

[27] Kolenikov, S. and Angeles, G. (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *The Review of Income and Wealth*, 55(1):128–165.

[28] Mazziotta, M. and Pareto, A. (2019). Use and misuse of pca for measuring well-being. *Social Indicators Research*, 142(2):451–476.

[29] McBride, L. and Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *World Bank Economic Review*, 32(3):531–550.

[30] Nadler, B. and Coifman, R. R. (2004). Partial least squares, Beer's law and the net analyte signal: statistical modeling and analysis. *Journal of Chemometrics*, 19:435–54.

[31] Pampalon, R., Raymond, G., et al. (2000). A deprivation index for health and welfare planning in quebec. *Chronic Dis Can*, 21(3):104–113.

[32] Poirier, M. J., Grépin, K. A., and Grignon, M. (2020). Approaches and alternatives to the wealth index to measure socioeconomic status using survey data: a critical interpretive synthesis. *Social Indicators Research*, 148(1):1–46.

[33] Russolillo, G. (2012). Non-Metric Partial Least Squares. *Electronic Journal of Statistics*, 6(none):1641 – 1669.

[34] Santeramo, F. G. (2015). On the composite indicators for food security: Decisions matter! *Food Reviews International*, 31(1):63–73.

[35] Seeman, T., Merkin, S. S., Crimmins, E., Koretz, B., Charette, S., and Karlamangla, A. (2008). Education, income and ethnic differences in cumulative biological risk profiles in a national sample of us adults: Nhanes iii (1988–1994). *Social Science and Medicine*, 66(1):72–87.

[36] Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer, New York, NY.

[37] Smith, J. P. (2004). Unraveling the ses: health connection. *Population and development review*, 30:108–132.

[38] Unal, I. (2017). Defining an optimal cut-point value in ROC analysis: An alternative approach. *Computational and mathematical methods in medicine*, 2017:1–14.

[39] Vincent, K. and Sutherland, J. M. (2013). A review of methods for deriving an index for socioeconomic status in british columbia. *Vancouver: Centre for Health Services and Policy Research*.

[40] World Bank (2015). *The state of social safety nets 2015*. The World Bank.

[41] Yoon, J. and Klasen, S. (2018). An application of partial least squares to the construction of the social institutions and gender index (sigi) and the corruption perception index (cpi). *Social Indicators Research*, 138(1):61–88.

[42] Yoon, J. and Krivobokova, T. (2018). Treatments of non-metric variables in partial least squares and principal component analysis. *Journal of Applied Statistics*, 45(6):971–987.