Universidad de San Andrés
Economics Department
MSc. in Economics

How Multidimensional is Welfare?
A Sparse Principal Components Analysis

Wendy BRAU

Advisor: Walter Sosa-Escudero

Ciudad Autónoma de Buenos Aires
December 10, 2021

Abstract

This paper attempts to measure the dimensionality of welfare by using Principal Component Analysis (PCA) with sparse loadings –which combines PCA with regularization techniques–, and uses nonlinear PCA techniques to handle mixed type data. Assuming that welfare can be represented by a subspace of a given data set, the hypothesis of multidimensionality of welfare states that more than one interpretable dimension is necessary to describe it. An empirical application to Argentina's Permanent Household Survey shows the limitations of PCA, and the advantages of PCA with sparse loadings, in determining the relevant subset of variables for assessing welfare. I find such a subset among 126 mixed type variables. I conclude that welfare is multidimensional, but there is room for dimensionality reduction: with three sparse principal components, it is possible to explain 20% of the variance using only 35% of the variables, and 30% of the variance using half of them. With a single sparse principal component, it is possible to explain 20% of the variability in welfare using half of the variables. This finding could be useful for implementing shorter household surveys.

Keywords: Welfare, Principal Component Analysis, Sparse PCA, nonlinear PCA, Regularization, Household Surveys, Argentina

Códigos JEL: C38, I31, C55

Contents

1 Introduction

Although there is an agreement in the literature regarding the multidimensional nature of welfare (Sen, 1985; Kakwani & Silber, 2008; Aaberge & Brandolini, 2015), in practice it is necessary to know *how many* and *which* variables should be considered to measure and describe it. Selecting the minimum set of features associated to welfare would allow to design shorter, easier to implement, and less expensive surveys, and to achieve higher response rates (Edo, Sosa-Escudero & Svarc, 2020).

Assuming that welfare can be measured from a dataset of K features, the multidimensionality hypothesis poses two related but different questions. The first is whether welfare can be summarized in one unique dimension, or how many dimensions are necessary. Formally, if we can project the initial dataset onto a space of dimension $P < K$ that captures its variability adequately, which is the value of P. The second question is whether this smaller set of dimensions has any meaningful interpretation. In other words, whether it can be associated to a specific aspect of welfare, that is, to a subset of the original variables –for example, income variables, or employment variables. In summary, the dimensionality of welfare refers to the number of interpretable dimensions needed to measure it. We would say it is unidimensional (i) if we could project it onto a space of one dimension that captures its variability adequately, and (ii) if that dimension consisted of one unique meaningful type of variables (e.g., income variables).

Principal Component Analysis (PCA) is useful for answering the first question. PCA is a dimensionality reduction technique that attempts to summarize the variability of a dataset of K features in $P < K$ principal components. The principal components are the orthogonal directions of maximum variability in the dataset, in decreasing order, obtained as a linear combination of the original variables. The subset of P components that accounts for a good portion of the variability generates a subspace of lower dimension that is an adequate representation of the original data. Therefore, a linear projection using PCA of a dataset that measures welfare can be useful to summarize welfare in fewer dimensions.

However, PCA has a disadvantage: in general, each principal component is a linear combination of *all* the original variables. To begin with, this makes interpretation difficult, as we cannot associate each component to a specific aspect of welfare. Moreover, if all variables should be used to obtain the components, it is not possible to find a subset of variables that capture welfare adequately, and thus, that can be used to design shorter surveys. Instead, the PCA with sparse loadings methods by Zou, Hastie & Tibshirani (2006) and Witten, Tibshirani & Hastie (2009) embed feature selection techniques in PCA to get each principal component as a linear combination of only a subset of original variables. Therefore, these methods are useful for answering the second question on dimensionality of welfare. That is, for finding interpretable components, associated to specific aspects of welfare. If we use only some of the components with sparse loadings, we can reduce the dimensionality and find a relevant subset of $K^* \subset K$ original features that captures the variability in welfare.

Assuming that welfare is adequately represented in a dataset with K features, the goal of this thesis is to select the relevant subset $K^* \subset K$ that accounts for its variability. For so doing, it explores the use of PCA with sparse loadings (Zou, Hastie & Tibshirani, 2006; Witten, Tibshirani & Hastie, 2009). An empirical use case with Argentina's Household Survey data called *Encuesta Permanente de Hogares* (Permanent Household Survey, EPH from now on) shows the advantages of this methodology. It must be noted that, even though the EPH contains some monetary and non-monetary *objective* aspects of welfare, there are relevant *subjective* factors that are not measured in the survey (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013). That is, the EPH does not cover all the relevant dimensions of welfare. However, finding the dataset that should be used as a starting point is out of the scope of this thesis. Therefore, all results are relative to surveys similar to the EPH, and they are useful to find relevant variables to make them shorter, although they will just cover the monetary and non-monetary aspects of welfare.

In turn, the EPH has mixed type variables: although there are some numerical ones, most are categorical, nominal and ordinary. As PCA and PCA with sparse loadings are designed for numerical variables, another contribution of this thesis is to explore the use of non-linear PCA techniques –also called PCA with optional scaling (Mori, Koruda & Makino, 2016)– to quantify categorical data. This is, to assign a number to each of the categories). In consequence, I can work with more initial variables and avoid an ad-hoc pre-selection. Using microdata of the EPHs for the third and fourth quarters of 2019 in the Gran Buenos Aires region, I construct two databases at the person level, with some variables referred to each person's household and dwelling. Each dataset has 126 variables, optimally scaled. Then, I implement PCA and different PCA with sparse loading models. The models are different in terms of (i) the methodologies used to obtain them, trying Sparse Principal Component Analysis (SPCA) by Zou, Hastie & Tibshirani (2006) and two specifications of Sparse Principal Components (SPC) by Witten, Tibshirani & Hastie (2009); (ii) the sparsity hyperparameters. Finally, an analysis of the stability of results over time allows to validate the methodology.

Next, the literature section discusses the differences and the advantages of the proposed methods with respect to previous studies. The third section describes the methodology and validation strategy. The fourth section describes the data and its processing, including optimal scaling. The fifth section presents three types of results: first, the criteria to select a model among the different estimated; second, a dimensionality of welfare analysis based in the selected model; third, the validation of the methodology. Conclusions at the end.

2 Literature

This thesis is part of the contributions of statistical and machine learning to the study of poverty, inequality and development. Specifically, it speaks to the literature on the multidimensionality of welfare. There is an agreement on the multidimensionality of welfare (Sen, 1985; Kakwani & Silber, 2008; Aaberge & Brandolini, 2015): it is not possible to measure it using only one meaningful dimension. However, in practice it is useful to know how many variables are relevant and should be included for adequately measuring it (Caruso, Sosa-Escudero & Svarc, 2015). Lacking this definition, welfare groups are often constructed based on income, which might be leaving out relevant welfare aspects. In addition, selecting the smallest set of variables needed to measure welfare would allow to designed shorter, easier to implement, and less expensive surveys, and to reduce the non-response rates.

Previous studies attempting to define the dimensionality of welfare have used unsupervised machine learning techniques like factor analysis and clustering. Most of them conclude that, although income is a relevant variable, welfare is indeed multidimensional and other aspects should be included. Gasparini, Sosa-Escudero, Marchionni & Olivieri (2013) use factor analysis over 12 welfare-related variables form the Gallup World Poll for Latin America. They define multidimensionality as the need of keeping more than one factor, and conclude that at least three factors are necessary. The first one is related to income; the second one, to subjective well-being; and the third one, to basic needs. Another study using factor analysis is the one by Ferro Luzzi, Flückiger, & Weber (2008), based on 32 variables from the Swiss Household

Panel, who conclude that four latent factors are needed, capturing financial, health-related, neighborhood, and social exclusion aspects.

Caruso, Sosa-Escudero & Svarc (2015) focus on the multidimensionality of poverty. They attempt to find a strict subset of the original variables that allows to identify the poor. They implement a two-step methodology. First, they use k-means to create clusters of poor and non-poor people, based on 15 variables from the Latin America *Gallup World Poll* –most of them, the ones used by Gasparini, Sosa-Escudero, Marchionni & Olivieri (2013). Second, the method for selection of variables for cluster analysis from Fraiman, Justel & Svarc (2008) allows them to select the smallest subset of variables needed to replicate the classification into poor and non-poor: household income, having enough money to buy food, and having a personal computer at home. Once again, income is a relevant variable, but not the only one.

Instead, Edo, Sosa-Escudero & Svarc (2021) center their attention on middle-class and the relevant variables that distinguish it from the rich and the poor. They use 19 household level variables from the 2004 to 2014 EPHs that are ordinal measures of welfare and that cover different aspects –like per capita household income, property and wealth, education and employment, housing characteristics, having a housemaid. Then, they construct an unidimensional welfare index, by using PCA and keeping the module of the first principal component to project data in a direction of increasing welfare. They define multivariate quantiles based on this index, establishing an upper and lower threshold that contains the middle-class. Finally, they also use Fraiman, Justel & Svarc (2008) to select the subset of variables that allows them to (i) reconstruct the unidimensional index with precision, (ii) distinguish the poor from the vulnerable middle-class, (iii) distinguish the vulnerable from the non-vulnebale middle-class, (iv) distinguish the non-vulnerable middle-class from the rich. They find that, on average, the most relevant variables to reconstruct the unidimensional index are paying in installments, income, occupation, and having a house maid. Thus, they conclude that welfare is multidimensional.

This thesis explores the suitability of PCA with sparse loadings techniques by Zou, Hastie & Tibshirani (2006) and Witten, Tibshirani & Hastie (2009) to perform unsupervised selection of the features that capture the variability in welfare. Tthe detailed description of these methods is in Section 3. The principal components from PCA are linear combinations of all original variables, which makes the interpretation of results difficult. For factor analysis, Gasparini, Sosa-Escudero, Marchionni & Olivieri (2013) and Ferro Luzzi, Flückiger, & Weber (2008) used rotations to get some variables with zero loadings in each component. However, rotations cannot guarantee that enough variables have zero loadings so as not to use all the original features to construct the principal components. Instead, PCA with sparse loadings linearly combines just a subset of the original variables in each component, which allows to find a strict subset K^*, and also have interpretable components. An alternative to rotations would be to establish ad-hoc thresholds to consider that the loadings of some variables below that threshold are in fact zero. However, Cadima & Jolliffe (1995) argue that the threshold approach might be unreliable, and simulations from Zou, Hastie & Tibshirani (2006) show that the explained variance is lower than when using PCA with sparse loadings.

As in Caruso, Sosa-Escudero & Svarc (2015) or Edo, Sosa-Escudero & Svarc (2021), the main goal is to select a strict subset of the original features. However, unlike the methodologies of those previous studies, PCA with sparse loading allows for unsupervised feature selection in just one step. It does not need to annotate data with a label to be predicted, like 'the poor' or 'the middle-class'. The selection of features will depend on the extent to which they explain the variability of welfare between people. PCA with sparse loadings has some of the interpretation advantages of PCA: the new dimensions are linear combinations of the original variables, and it is possible to quantify the loss in explained variance when excluding some of them. Finally, unlike clustering methods, an advantage of PCA with sparse loadings is that the fitted model can be implemented on a new dataset using the loadings matrix, without need of training the algorithm again for new datasets, as it defines a structure of variable relationships that can be applied to any new sample.

The second contribution of this thesis is to use optimal scaling of categorical variables. Optimal scaling techniques assign a numerical value to each of the categories by optimizing any parameter that serves to the goal of the final analysis while complying with the constraints of the variables to be transformed. In optimal scaling for PCA, the goal is to maximize the sum of the first eigenvalues of the correlation matrix of quantified data (Mair & De Leeuw, 2010). In turn, the constraint for ordinal variables is that the quantification preserves the original ordering of categories, though distances between them may vary (Mori, Koruda & Makino, 2016). This technique allows me to include in the analysis almost all the variables from the EPH. While the reviewed studies start with less than 40 variables, I start from 126, avoiding an ex-ante selection on the variables to consider.

Merola & Baulch (2018) is the closest study to this thesis, at it combines PCA with sparse loadings and PCA with optimal scaling for a poverty, inequality, and development analysis. However, their main goal is not related to the dimensionality of welfare, but to construct an index of assets, based on the ownership of 34 different assets recorded in household surveys from Vietnam and Laos. In fact, many studies that use machine learning to study poverty, inequality and development use PCA to construct asset indexes that then are used as the dependent variable of classification models (such as in Blumenstock, Cadamuro, & On, 2015, or Jean et al., 2016). Merola & Baulch (2018) use optimal scaling to preserve the ordering of the original asset counts when constructing the asset index. They found that the index that uses optimal scaling improves the prediction of income rankings at the household and per capita level. Additionally, they use PCA with sparse loadings to find which are the main assets that explain income variation between households. They find that durable foods, ownership of transport means, and housing characteristics are relevant, and that the indexes they constructed with sparse loadings –using between a half and a third of total assets– achieve results similar to PCA when predicting income.

3 Methods

Let X be a data matrix of dimension $N \times K$, where K is the number of variables in the EPH and N the number of people. The variables are centered and scaled to have unit variance. The goal is to find a relevant subset $K^* \subset K$ to explain the variability in X. That is, to perform unsupervised feature selection.

3.1 PCA

PCA is an unsupervised method to reduce data dimensionality. It linearly projects the original K variables onto a new space of K components, which are the directions of maximum variability in the data, sorted in decreasing order, and orthogonal to each other. Then, it is possible to select a subset of $P < K$ principal components that form a lower dimensional space that captures as much variability as possible. The P principal components can be thought of as latent factors that generate the variation in the data, where each of them summarizes a set of correlated original variables (James et al., 2013).

Let $F_{N \times K}$ be the matrix of components, where the first column F_1 is the first principal component and each row $i \in 1...N$ in F_1 is the value of the observation i in the first principal component. F is a linear combination –optimal in terms of explained variance– between X and a matrix of loadings $V_{K \times K}$:

$$F = XV$$

The matrix V has the loading of each original variable in each principal component, in other words, the weight of each variable in the linear combination that each component is. V is constructed in such a way that F_1 has the highest variability. If Σ is the covariance matrix of X and V_1^* the first column of V, then $Var(F_1) = {V_1^*}^T \Sigma V_1^*$. Therefore:

$$V_1^* = \underset{V_1}{\arg\max} \quad V_1^T \Sigma V_1 \quad \text{s.t.} \quad V_1^T V_1 = 1$$

Where V_1 are all the possible loading vectors. The remaining principal components F_k with $k = \{2, .., K\}$ are defined by imposing an orthogonality constraint:

$$\begin{aligned} V_k^* = \underset{V_k}{\arg\max} \quad & V_k^T \Sigma V_k \\ \text{s.t.} \quad & V_k^T V_k = 1 \\ & Cov(F_k, F_{k-1}) = 0 \end{aligned}$$

Alternatively, F can be obtained from the singular value decomposition of X. Let D be the diagonal matrix where $d_1, ...d_K$ in the diagonal are the eigenvalues of Σ from highest to lowest, and V the matrix with the corresponding eigenvectors $V_1^* ... V_K^*$, then:

$$X = UDV^T, \quad U^T U = I_N, \quad V^T V = I_K, \quad d_1 \geq d_2 \geq, ..., \geq d_K \geq 0 \quad \Rightarrow XV = UD = F \tag{1}$$

The variance of the j^{th} component is $Var(F_j) = {V_j^*}^T \Sigma V_j^* = d_j$. Since components are uncorrelated, the total variance explained by the P principal components is the sum of the eigenvalues $\sum_{j=1}^{P} d_j = Tr(D)$, and the proportion of variability that each component j explains is therefore $\frac{d_j}{Tr(D)}$.

3.2 Sparse loadings

The disadvantage of PCA is that each principal component is a linear combination of the K original variables, that is, all variables typically have non-zero loadings V (Zou, Hastie & Tibshirani, 2006). This hinders the interpretability of components and selecting a strict subset K^* from the original K features. Alhough PCA allows for dimensionality reduction meaning obtaining a new projected space of lower dimensionality, it does not allow for dimensionality reduction meaning feature selection.

To tackle this problem, PCA with sparse loading methods aim at increasing the number of variables with zero-loadings in each principal component. If some variables have zero loadins in the first principal component, there is a $P < K$ such that keeping the P principal components explains the highest possible variability with just a subset $K^* < K$ of the original features. This allows for unsupervised feature selection.

3.2.1 Regularization

The PCA with sparse loading methods use feature selection or regularization techniques LASSO (*Least Absolute Shrinkage and Selection Operator*) and Elastic Net. From a linear regression model with N observations and K predictors, where $Y_{N\times 1}$ is the response and $X_{N\times K}$ the matrix of predictors, both standardized, the LASSO estimator for the coefficient of the j^{th} predictor can be obtained as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} ||Y - \sum_{j=1}^{K} X_j \beta_j||^2 + \lambda \sum_{j=1}^{K} |\beta_j|$$

Where $\lambda \geq 0$ is a L1 norm penalty that, if large enough, leads to some of the estimated coefficients to be zero. In turn, the Elastic Net estimator adds and extra quadratic penalty $\lambda_2 \geq 0$:

$$\hat{\beta}_{EN} = (1+\lambda_2)\Big\{ \arg\min_{\beta} ||Y - \sum_{j=1}^{K} X_j \beta_j||^2 + \lambda_2 \sum_{j=1}^{K} \beta_j^2 + \lambda \sum_{j=1}^{K} |\beta_j| \Big\}$$

In both cases, sparsity in the consequence of the L1 norm penalty (λ), which leads to corner solutions when finding the vector of estimators. Instead, the quadratic penalty (λ_2) shrinks estimated coefficient *towards* zero, and it is used in Ridge estimators:

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} ||Y - \sum_{j=1}^{K} X_j \beta_j||^2 + \lambda_2 \sum_{j=1}^{K} \beta_j^2$$

3.2.2 SPCA and SPC

I will use two methods to find principal components with sparse loadings: *Sparse Principal Component Analysis* (SPCA) by Zou, Hastie & Tibshirani (2006) and *Sparse Principal Components* (SPC) by Witten, Tibshirani & Hastie (2009). Both methods are implemented in R libraries by the same authors, `elasticnet` and `PMA` respectively (Zou & Hastie, 2020; Witten & Tibshirani, 2020). Giménez (2015) reviews other methods previously used for finding sparse loadings.

Let $X_{N\times K}$ be the data matrix with N observations and K variables, $F_{N\times K}$ the components matrix and $V_{K\times K}$ the loadings matrix.

SPCA defines the vector of loadings in PCA as the result of a regression problem where the LASSO penalty can be introduced. LASSO renders sparse estimations, and therefore, sparse loadings. Zou, Hastie & Tibshirani (2006) prove the following main steps for implementing SPCA:

1. Let F_p be the p^{th} principal component and V_p the associated vector of loadings. Given that $F = XV$, it is possible to get V_p from F_p and X. Let $\lambda_2 \geq 0$ and

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} ||F_p - \sum_{i=1}^{k} X_j \beta_j||^2 + \lambda_2 \sum_{j=1}^{k} \beta_j^2 \tag{2}$$

then $V_p = \frac{\hat{\beta}_{Ridge}}{||\hat{\beta}_{Ridge}||}$

2. Adding the LASSO penalty in (2), $\lambda \geq 0$,

$$\hat{\beta} = \arg\min_{\beta} ||F_p - \sum_{j=1}^{K} X_j \beta_j||^2 + \lambda_2 \sum_{j=1}^{K} \beta_j^2 + \lambda \sum_{j=1}^{K} |\beta_j| \tag{3}$$

   V_p can be approximated as $\hat{V}_p = \frac{\hat{\beta}}{||\hat{\beta}||}$. If λ is large enough, $\hat{\beta}$ is sparse, and therefore the loadings vector V_p is sparse.

3. However, (3) requires knowing the principal components F to get the loadings matrix V. The problem can be modified to get both simultaneously, resulting in what the authors call the "SPCA criteria" to obtain the first P principal components. Let x_i be the i^{th} row vector of matrix X. Let $A_{K\times P} = [\alpha_1, ..., \alpha_P]$, $B_{K\times P} = [\beta_1, ..., \beta_P]$ and $\lambda_p, \lambda_2 \geq 0$, then

$$(\hat{A}, \hat{B}) = \arg\min_{A,B} \sum_{i=1}^{N} ||x_i - AB^T x_i||^2 + \lambda_2 \sum_{p=1}^{P} \sum_{j=1}^{K} \beta_{jp}^2 + \lambda_p \sum_{p=1}^{P} \sum_{j=1}^{K} |\beta_{jp}| \quad \text{s.t.} \quad A^T A = I_{P\times P} \tag{4}$$

   so $\hat{\beta}_p \propto \hat{V}_p$ for $p = 1, ..., P$. The greater λ_p for the p^{th} principal component, the sparser its loadings vector.

The authors propose an iterative algorithm to solve (4), which requires defining the sparsity parameter λ_p. For so doing, they suggest trying different λ_p and choosing that which is a good compromise between explained variance and sparsity.

Unlike PCA, SPCA does not impose orthogonal loadings, and therefore the components can be correlated. If F_p and F_{p-1} are correlated, the variance explained by component F_p may reflect in part the contributions from $F_1, ..., F_{p-1}$. Therefore, the authors propose a formula to adjust the calculation of the total variance explained by the first p components, using the lineal projection of F_p in the orthogonal subspace generated by $F_1, ..., F_{p-1}$. That is, linearly regressing F_p in the preceding components, and using the residual of that regression when calculating the explained variance.

SPC is a particular case of a penalized matrix decomposition method by Witten, Tibshirani & Hastie (2009), who also designed an iterative algorithm for implementing it. Let u be a column vector from U in the signular value decomposition of X in (1), v a column vector from V and $||w||_p$ the L_p-norm of vector w, that is, $(\sum_i w_i^p)^{\frac{1}{p}}$. SPC can be obtained as:

$$(u, v) = \arg\max_{u,v} u^T X v \quad \text{s.t.} \quad ||v||_1 \leq c, \quad ||u||_2^2 \leq 1, \quad ||v||_2^2 \leq 1 \tag{5}$$

Where v is the vector of loadings of the component and c the sparsity parameter: it sets an upper bound to the sum of absolute values of the loadings of v. The smaller c, the sparser v. The authors propose a strategy similar to cross validation to choose the value of this parameter. Briefly, it consists on the following steps. For a grid of possible values for c, construct 10 new matrices based on X, called $X_1, ...X_{10}$, by removing a random disjoint subset with 10% of observations from each. For each $\hat{X}_m$ (with $m \in 1, ...10$), estimate the first principal component via SPC, compute $\hat{X}_m = U_m D_m V_m$ and calculate the mean squared error of estimating X_m using $\hat{X}_m$. Then, get the average of the mean squared errors for $X_1, ...X_{10}$, and choose the parameter c that renders the lowest value. While this is a criterion based on the stability of estimations, the authors point out that if instead the goal is to better understand the data structure, other criteria might be preferable, like achieving certain desired sparsity levels.

Like SPCA, this method does not guarantee orthogonality between v_k and v_{k-1}. Therefore, they propose another method to get an approximation to orthogonality, solving for $k > 1$:

$$(u_k, v_k) = \underset{u_k, v_k}{\arg\max}\, u_k^T X v_k \quad \text{s.t.} \quad ||v_k||_1 \leq c, \quad ||u_k||_2^2 \leq 1, \quad ||v_k||_2^2 \leq 1, \quad u_k \perp u_1, ..., u_{k-1} \tag{6}$$

3.3 Optimal Scaling

PCA is a technique designed for numerical variables, while the EPH is a dataset with mixed types: it has numerical variables, but most are categorical, both nominal and ordinal. In general, previous studies based on these data used the ad-hoc coding from the survey as numerical values, or transformed categorical variables in binary variables, which introduce unnecessary constraints:

- They impose the same distance between adjacent categories, or between each category and the base category.
- For nominal variables, they impose a totally arbitrary increasing order among categories.
- For ordinal variables, when binary variables are used, the original ordering among categories other than the base is lost (Merola & Baulch, 2018).

Instead, optimal scaling techniques transform categorical variables into numerical variables imposing the minimum constraints necessary, deciding the quantification (i.e., the number associated to each category) so it optimizes some criterion associated to the goal of the analysis. In the case of PCA, so it maximizes the first P eigenvalues of the correlation matrix of quantified data (Mair & De Leeuw, 2010).

Let y_j a qualitative vector with W_j categories and N observations to quantify. First, y_j is coded using an indicator matrix $G_{N \times W_j}$ with binary column vectors. Let $y_j^* = G_j q_j$ be the optimal scaled vector, then the goal is to find q_j, the optimal scaling, subject to the following constraints depending on the type of y_j:

- Nominal: the quantification is unconstrained, only guaranteeing that observations from the same category get the same numerical value.
- Ordinal: the quantification is subject to preserving the ordering of the categories. If categories w_1 and w_2 are such that $y_{jw_1} < y_{jw_2}$, then the quantified categories should have the same order, $y^*_{jw_1} < y^*_{jw_2}$.
- Numerical: already quantified, y_j is simply standardized.

PCA with optimal scaling is also called 'nonlinear PCA' (Mori, Koruda & Makino, 2016). I implement optimal scaling for categorical variables using the R library `aspect` by Mair & De Leeuw (2018).

3.4 Estimation

I estimate 9 models: PCA, two versions of SPCA by Zou, Hastie & Tibshirani (2006) implemented in the R library `elasticnet`, and six versions of SPC by Witten, Tibshirani & Hastie (2009) implemented in the R library `PMA`.

1. `PCA`
2. `SPCA varnum`: I use an ad-hoc restriction so the n^{th} principal component has at most $16 - n$ variables with non-zero loadings, with $n \in 1, 2, 3, ..., 15$. That is, the first principal component will have 15 variables with non-zero loadings.
3. `SPCA lambda`: I choose the sparsity parameter λ that achieves a good balance between explained variance and sparsity for the first components. As explained in Section 5.1.2, I choose $\lambda = 0.1$.
4. `SPC`: estimated with the methodology that does *not* impose orthogonality from equation (5), and a sparsity parameter $c = 6.4$, chosen with the cross validation method suggested by Witten, Tibshirani & Hastie (2009).
5. `SPC sparse`: estimated with the methodology that does *not* impose orthogonality from equation (5), and a sparsity parameter $c = 3.4$ which prioritizes a greater sparsity, as explained in Section 5.1.2.
6. `SPC orthog.`: estimated with the methodology for obtaining approximately orthogonal components from equation (6), and a sparsity parameter $c = 6.4$, chosen with the cross validation method suggested by Witten, Tibshirani & Hastie (2009).
7. `SPC orthog. sparse`: estimated with the methodology for obtaining approximately orthogonal components from equation (6), and a sparsity parameter $c = 3.4$ which prioritizes a greater sparsity, as explained in Section 5.1.2.
8. `SPC pos. orthog.`: estimated with the methodology for obtaining approximately orthogonal components from equation (6), with a sparsity parameter $c = 5.2$, and adding the constraint that loadings should be positive.
9. `SPC pos. orthog. sparse`: estimated with the methodology for obtaining approximately orthogonal components from equation (6), a sparsity parameter $c = 3.2$ which prioritizes a greater sparsity, and adding the constraint that loadings should be positive.

3.5 Validating the methodology

If welfare is stable over time -which is more likely the closer the periods being compared-, a way to assess the performance of the methodology is to see if its results are stable over time.

Firstly, as Edo, Sosa-Escudero & Svarc (2020) do, I verify whether the loadings for each variable in the first principal components are stable between surveys at different but close points in time.

Secondly, let $F_{N \times P}$ be the matrix of the values for each person in each of the P principal components. In other words, F is the matrix of the locations of each individual in welfare measures constructed with P principal components. It is possible to get F from the loading matrix V and the original data matrix X via $F = XV$. This means that there could be two ways of getting the matrix of locations in t, F_t: either by using the loadings matrix from implementing any of the PCA methods over data from t (V_t), or by using the loading matrix obtained from data in $t-1$ (V_{t-1}). Then, I compare:

- Real location $F_t = X_t V_t$
- Predicted location $\hat{F}_t = X_t V_{t-1}$

On the one hand, I graphically compare the distributions of F_t and $\hat{F_t}$ values for each component. On the other hand, I construct error metrics. If F_p is the p^{th} principal component, I can get the Mean Squared Error (MSE) of predictions as:

$$MSE_t = \frac{\sum_{p=1}^{P} \sum_{i=1}^{N} (\hat{F}_{ipt} - F_{1ipt})^2}{N + P}$$

That being said, MSE_t compares the *absolute* locations of each individual in the welfare space. Instead, it can be more interesting to compare the *relative* location of individuals. Therefore, I measure and compare the *relative* locations on individuals in the welfare space F_t and $\hat{F_t}$. There are two cases that can arise when attempting to measure the relative location: that only the first principal component was used for measuring welfare ($F_{N\times 1}$ and $\hat{F}_{N\times 1}$, the unidimensional case), or that $P > 1$ principal components were used ($F_{N\times P}$ and $\hat{F}_{N\times P}$, the multidimensional case).

3.5.1 Unidimensional relative location (first principal component of welfare)

In the unidimensional case, each individual can be located in the quantiles of F_t and in the quantiles of $\hat{F_t}$. Then, I can calculare the mean squared error of the classification in real quantiles and predicted quantiles ($MSEQ$), the percentage of misclassified people (the complement of accuracy, that I will call *inacc*), and the maximum distance between the right and the wrong quantile for each individual (*maxdist*). Let Q_{it} be the quantile for person i in F_t and $\hat{Q}_{it}$ the quantile for person i in $\hat{F_t}$, the mentioned metrics are calculated as:

$$MSEQ_t = \sum_{i=1}^{N} \frac{(\hat{Q_{it}} - Q_{it})^2}{N}$$

$$inacc_t = \sum_{i=1}^{N} \frac{\mathbb{1}[|MSEQ_{it}| > 0]}{N} \cdot 100$$

$$maxdist_t = max\{\ |\hat{Q_{it}} - Q_{it}|, \quad i \in 1, ..., N\ \}.$$

The more quantiles considered, the more strict the criterion. At one extreme, the ordering between individuals would be required to be exactly the same in F and $\hat{F}$. I will consider quartiles and deciles.

3.5.2 Multidimensional relative location ($P > 1$ principal components of welfare)

To calculate relative locations in a space of $P > 1$ principal components of welfare like $F_{N\times P}$ and $\hat{F}_{N\times P}$, an option is to use multivariate quantiles. I calculate depth quantiles for each individuals. It must be noted that defining a quantile of central depth resonates with the middle class definition adopted by Gigliarano & Mosler (2009). In particular, I consider two depth measures:

1. Tukey's depth. The depth of point i is the minimum number of points of any semispace containing i.
2. Mahalanobis' distance to the spatial median. Let i_M be the spatial median, the point that minimizes the sum of absolute distances to the rest of the observations. The Mahalanobis' distance from point i to i_M is $\sqrt{(i - i_M)^T \Sigma^{-1} (I - I_M)}$, where Σ is the covariances matrix of X.

After locating each individual in a depth quantile, I can construct the same error metrics than in the unidimensional case: $MSEQ_t$, $inacc_t$ and $maxdist_t$.

Depending on the spatial data distribution, a measure of depth around a center may not be very meaningful for identifying groups of individuals in close proximity to each other. In such cases, another option is to use a clustering method, to locate each individual in a cluster from F_t and a cluster of $\hat{F}_t$, and to analize the correspondence between the two.

4 Data

The Argentinian National Institute of Statistics and Census (INDEC) has the microdata of the Permanent Household Survey (EPH) publicly available. I obtained these data for the Gran Buenos Aires region in the third and fourth quarters of 2019, as well as the National Classification of Ocupations (CNO) and the Economic Activities Classification for Sociodemographic Surveys (CAES) from the R package `eph` (Kozlowski et al., 2020). I used the data of the fourth quarter for selecting the modeling and analyzing the dimensionality of welfare, and the data of the third quarter for validating the methodology. The preprocessing detailed in 4.1 resulted in two databases with 126 variables each, 16525 individuals for the third quarter, and 15610 individuals for the fourth quarter. There are different types of variables and each was optimally scaled, as explained in 4.2 and 4.3.

4.1 Preprocessing

These were the preprocessing decisions:

- *Level.* Combining person-level and household-level surveys to have a unique database at the person-level, but including the household characteristics of each person.

- *Ordering based on welfare.* The EPH has variables that can be ordered from lower to higher welfare (for example, income), while others that cannot (for example, the type of economic activity). Whenever possible, I sorted the categories so higher values represent higher welfare.

- *New variables.* I used many of the original variables in the EPH but I also constructed additional ones based on them, such as the variables in Edo, Sosa-Escudero & Svarc (2020), and those that are part of the Household Living Conditions Indicators by INDEC (Gómez et al., 2004; INDEC, 2019).

- *Non-responses* resulting from people who did not want to answer the question they were asked were coded differently than those missing because of the hierarchical structure of the questionnaire (for instance, there is a set of questions for unemployed people only). Then, I removed observations with the first type of non-response. This means that the analysis will only be representative of those people who answer all the questions they are asked (approximately 55% of the people, with most non-responses income variables). In future work, I would like to explore imputation methods.

- I deleted variables with zero variance and those perfectly correlated with others, and unified variables whenever possible (e.g., when one question further splits into categories one of the categories of the previous question). I also removed a few observations with inconsistent answers to different but related questions.

4.2 Types of variables

To facilitate the analysis, note that the variables can fall into four topics: *Income*, *Employment*, *Housing* and *Household Members*. The latter includes some sociodemographic features like health and education. Moreover, variables have different levels (person, household, dwelling), and some can be sorted from lower to higher welfare, while others cannot. These details, alongside a brief description of each variable, can be found in Annex A. Many variables are defined in the same way as INDEC (2020).

4.3 Optimal scaling

To assign numerical values to categorical variables I used optimal scaling, as described in Section 3.3. The quantification maximizes the first eigenvalue of the correlation matrix of quantified data, while it imposes constraints depending on the type of variable:

- *Nominal variables.* As there is no constraint for the scaling of these variables, the order and distances between the values for coding each category might change. For example, the original categories of the variable ESTADO were 1 for employed, 2 for unemployed, and 3 for inactive people. After the scaling, the values are -1.112 for employed, 0.905 for unemployed and 0.899 for inactive people. That is, employed and unemployed are the most dissimilar.
- *Ordinal variables.* The scaling is constrained to preserve the original ordering between the categories, but the distance between them may change. For example, the variable about the quality of housing materials had original categories 0 for 'insufficient', 1 for 'partially insufficient' and 2 for 'sufficient', while the quantification has the values -2, -1.2 and 0.6 respectively. That is, the distance from 'partially insufficient' to 'sufficient' is smaller than to 'insufficient'.
- *Numerical variables* are simply standardized.

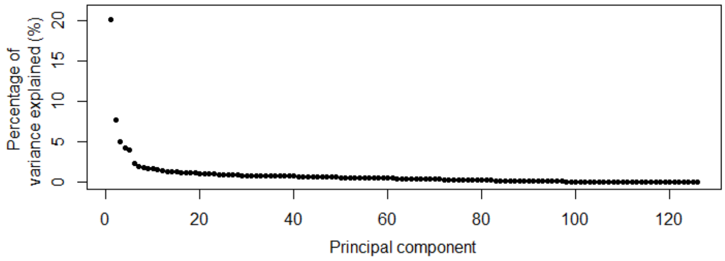
5 Results

To begin with, I will compare the estimated models (listed in 3.4) and I will explain the criterion for selecting two of the sparse models for the rest of the analysis: `SPC orthog.` and `SPC orthog. sparse`. Then, I will analyze the dimensionality of welfare based on those two models. Finally, I will validate the methodology by analyzing the stability of the results over time.

5.1 Selecting models with sparse loadings

This subsection shows the results from PCA and PCA with sparse loading models in terms of the variance explained and the sparsity –the number of variables with zero loadings– of the first components. In general, there is a trade-off between sparsity and variance explained. However, each component from `SPC orthog.` explains almost the same amount of variability than each from `PCA` with significant sparsity gains: half of the variables would suffice for getting the first sparse principal component of the former, while all the variables are needed in the latter. In turn, `SPC orthog. sparse` allows for even greater sparsity gains if three principal components are used to explain the same variability than the first principal component from `PCA`: just 35% of the original variables need to be included. In other words, 20% of the variability

Figure 1: Variance explained by each component of `PCA`



of welfare can be captured either by the first sparse principal component from `SPC orthog.`, that can be constructed with half of the original features, or by the three principal components from `SPC orthog. sparse`, that can be constructed with just 35% of the original features.

5.1.1 PCA

Figure 1 shows the variance explained by each of the principal component of `PCA`. The first component explains 20% of total variance, a significant amount given that the starting point are 126 variables, and noticeably more than the remaining components (25 times the average variance explained by each component). As a reference, Edo, Sosa-Escudero & Svarc (2020) use the first component of a PCA from 19 variables that explains 30% of the variance as a welfare index, and Merola & Baulch (2018) use a first component that explains 21% of the variance of 34 variables as an asset index. Then, while components from second to fifth explain between 7% and 4% of variance, the pattern changes since the sixth principal component, that explains the 2%.

Figure 2 shows that all variables have non-zero loadings in the first two principal components from `PCA`, which makes interpretation difficult. It is true that there is a pattern regarding which variables have smaller or greater loadings. In the first component, the variable ESTADO has a loading much more negative than the others, and the variables with greater loadings are the deciles of the income of the main occupation. In turn, in the second component the variables with greater loadings are related to household income. However, it would not be possible to select a subset K^* of the features to get the first component without having to define *ad hoc* thresholds under which variables will be assigned zero loadings, and this strategy is not trustworthy (see Section 2). Instead, using sparse methods may allow for interpretable components, related to a subset of the original variables only, and to find a minimum subset of K^* features that accounts for a good portion of the variability in welfare.

5.1.2 Selecting the sparsity hyperparameter

When implementing some of the SPCA and SPC models, I should choose the sparsity hyperparameter: λ in SPCA and c in SPC. Figure 3 plots the explained variance (vertical axis) against the sparsity (horizontal axis) for each of the three principal components (columns), under different values of the sparsity hyperparameter (the number next to each point), and for three different models (rows): `SPCA lambda`, `SPC sparse` and `SPC orthog. sparse`. There is a trade-off between explained variance and sparsity: hyperparameters that render greater sparsity result in lower variance explained by each component.

The intensity of such trade-off speaks of the multidimensionality of welfare. The more pronounced, it means

Figure 2: Loadings of each variable in the two first principal components of PCA

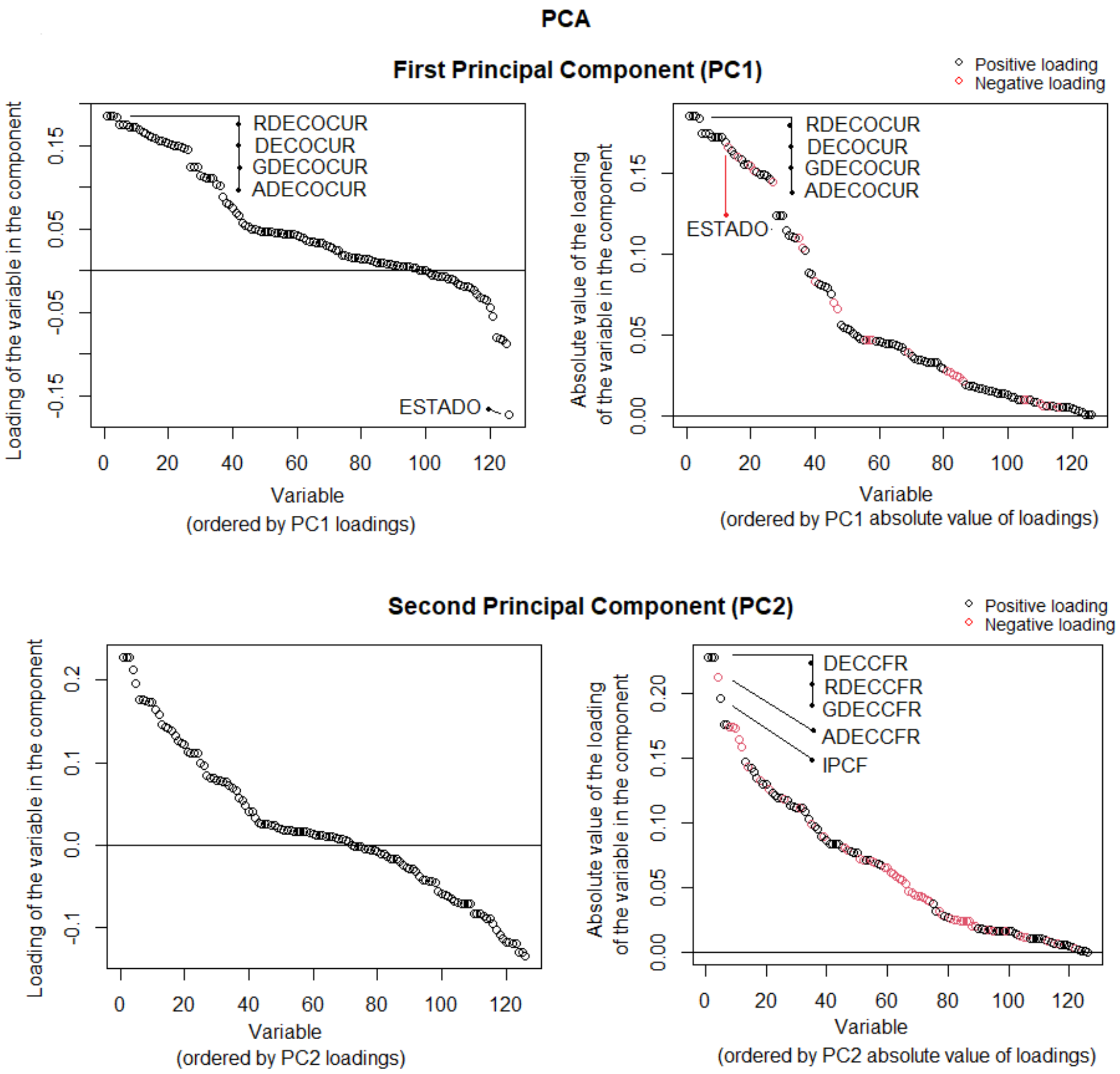
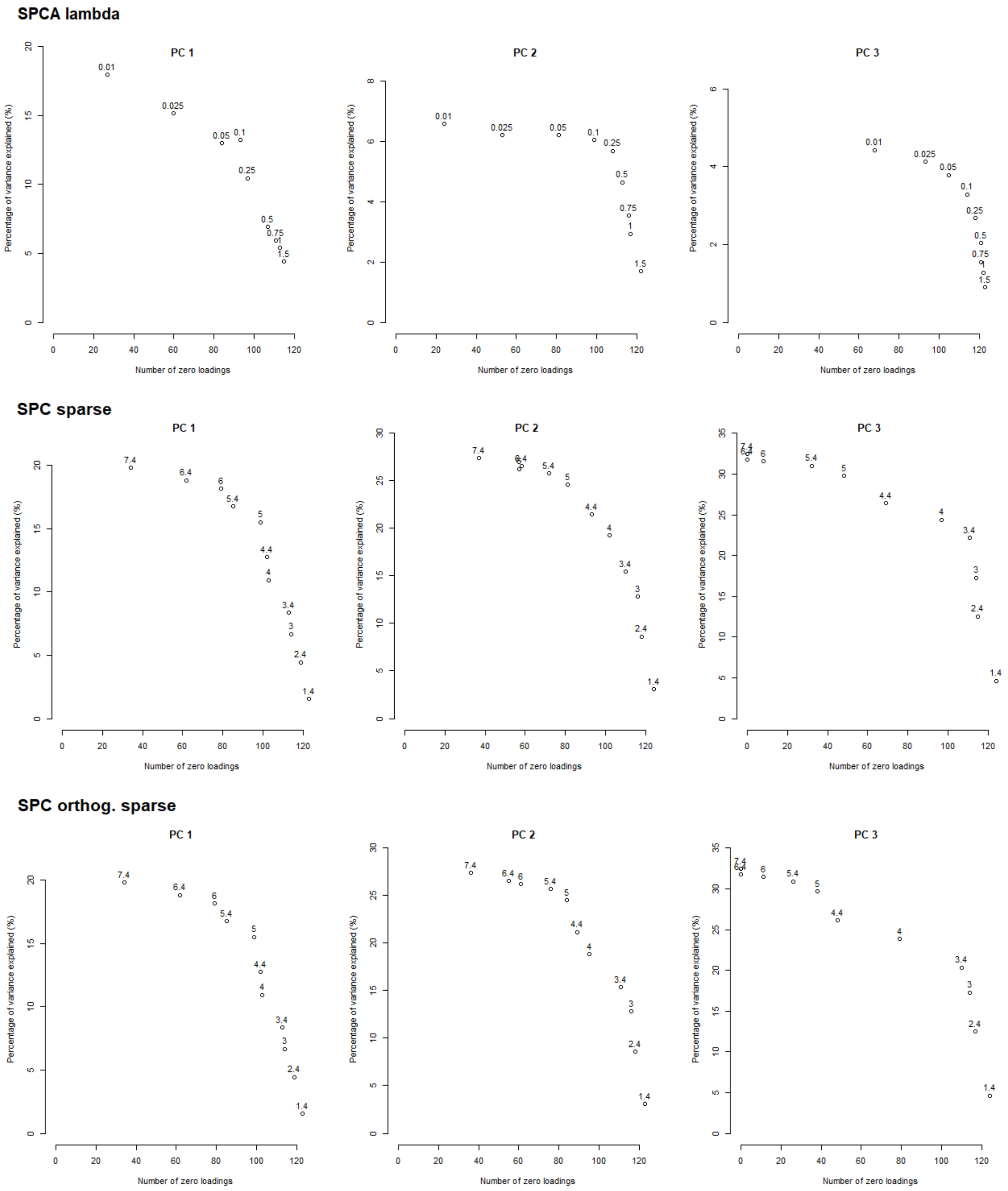


Figure 3: Explained variance versus sparsity trade-off when choosing the sparsity hyperparameter, by model



Note: the numbers next to each point indicate the value of the sparsity hyperparameter.

that more original features are needed to explain a significant portion of the variability of welfare, and therefore, that welfare is 'more multidimensional'. If welfare was unidimensional, there should be a point in the upper-right corner: most of the variance would be explained with just one of the original features, and there wouldn't be an explained variance versus sparsity trade-off. In turn, a more pronounced trade-off would be represented by a diagonal line from the upper-left corner to the lower-right corner: any sparsity increase would proportionally decrease the explained variance. Figure 3 shows that there is a trade-off, which suggests that welfare is multidimensional, but that there is room for dimensionality reduction: the points that are closer to the upper-right corner represent hyperparameters that allow for greater sparsity gains with lower explained variance losses.

Following Zou, Hastie & Tibshirani (2006), I choose a value for the sparsity hyperparameter that has a good balance between explained variance and sparsity. For `SPCA lambda`, $\lambda = 0.1$, because it increases the sparsity of the first component with even more explained variance than $\lambda = 0.05$, and it is located in the upper-right zone of the second and third components' plots. For the SPC models, the decision should be between $c = 5$ and $c = 3.4$, depending on the preferences between sparsity and variance. As I also implemented the `SPC` and `SPC orthog.` versions of SPC which prioritize variance ($c = 6.4$, see Section 3.4), I choose $c = 3.4$ to have another couple of models which prioritize sparsity instead.

5.1.3 Comparing models

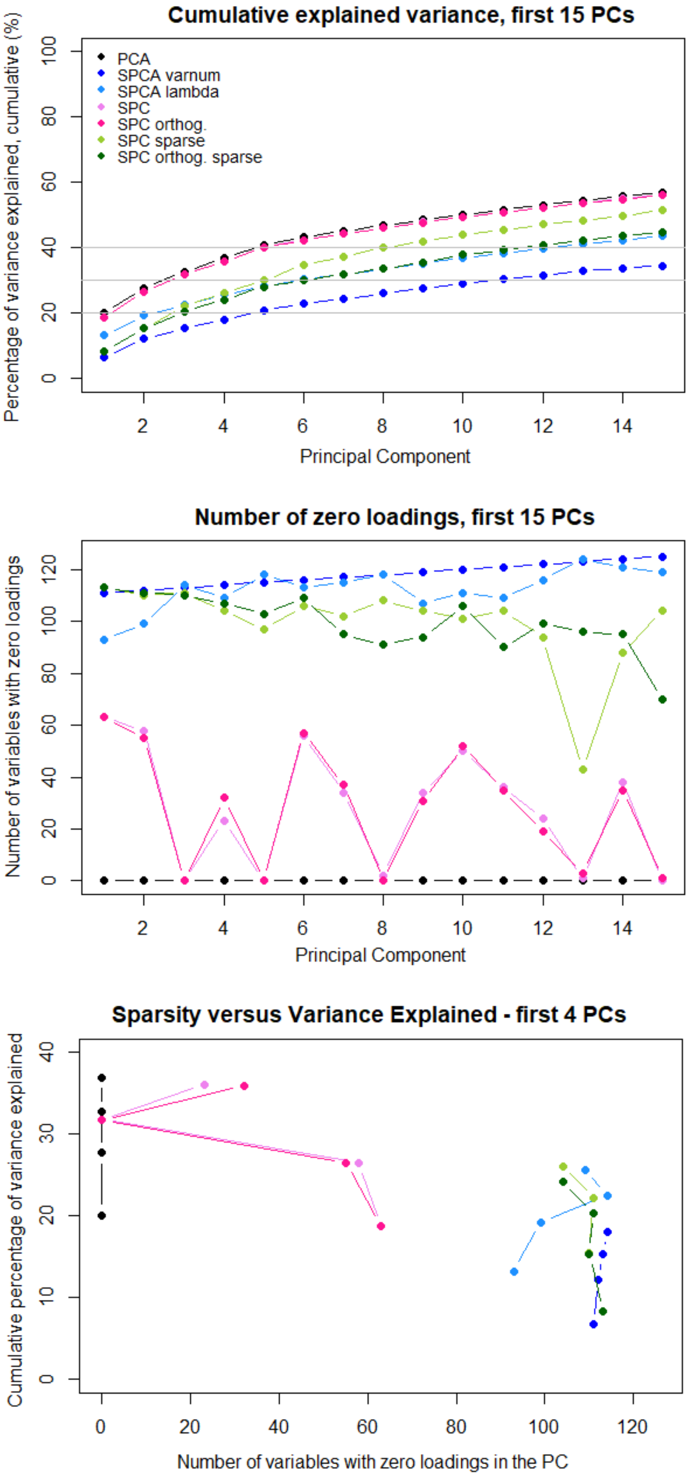
Figure 4 compares the explained variance and sparsity of each model by looking at the first 15 principal components. The middle panel shows that `PCA` is not useful at all to find a subset of K^* features that summarise the variability of welfare: no matter how much variability we may want to capture (and therefore how many principal components we choose), we always need to use all the variables to get the principal components. Therefore, a sparse model is needed, but which one is better?

There is also an explained variance versus sparsity trade-off when choosing between models, which again suggests that welfare is multidimensional. There are two main groups of models. On the one hand, `SPC` and `SPC orthog.` prioritize explained variance. The first components capture almost the same variance than `PCA` (upper panel), but they have less variables with zero loadings than the rest of the models (central panel). That said, the fact that the first component of these models captures the same variance than `PCA` does, but half of the variables in it have zero loadings, means that there is room for dimensionality reduction. On the other hand, SPCA and the sparse versions of SPC prioritize sparsity at the expense of explained variance.

Among SPC models, the approximately orthogonal versions are similar to the non-orthogonal in terms of variance and sparsity of the first five components. Therefore, I keep approximately orthogonal versions because they have interpretation advantages: they allow for biplots, correlation analysis, and measuring distances between people in spaces defined by many principal components without great distortions. Among SPCA models, I discard `SPCA varnum` because it explains little variance compared to the rest.

In summary, this leaves three pre-selected models: `SPC orthog.`, `SPCA lambda` and `SPC orthog. sparse`. Table 1 shows, for `PCA` and the selected models, how many principal components and how many features $K^* \subset K$ are needed for capturing a given amount of variance. This reveals another way of thinking about the tradeoff between sparsity and explained variance: for a desired level of explained variance, there is a trade-off between sparsity and how many components we need to use. In other words, there is a trade-off between lowering the number of the original variables (how many questions to ask in a survey) and

Figure 4: Comparing the explained variance and sparsity of each model



Cumulative explained variance, first 15 PCs

Number of zero loadings, first 15 PCs

Sparsity versus Variance Explained - first 4 PCs

Table 1: Comparing models in terms of number of principal components (PCs) required and sparsity achieved for different levels of explained variance

Perc.	PCA		SPC orthog.		SPCA lambda		SPC orthog. sparse	
variance	#	#	#	#	#	#	#	#
explained	PCs	variables	PCs	variables	PCs	variables	PCs	variables
8%							1	-113
13%					1	-93	2	-98
20%	1	all	1	-63	2	-71	3	-83
25%	2	all	2	-37	4	-59	4	-64
30%	3	all	3	all	6	-46	6	-50
40%	5	all	5	all	12	-19	12	-11

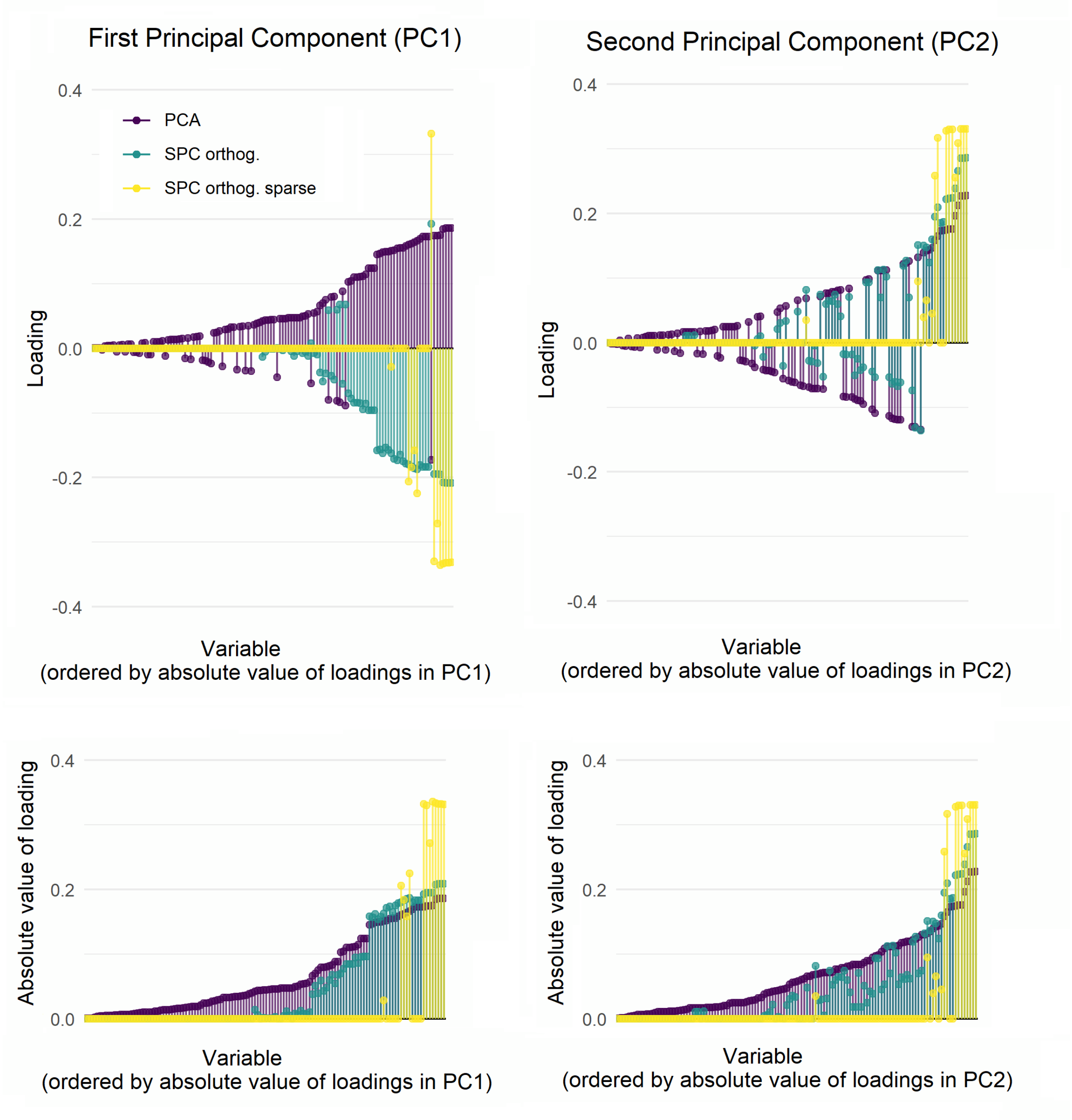
lowering the number of new dimensions (how many principal components to use in the analysis). The three preselected models occupy different places in this trade-off. I discard `SPCA lambda` because it is in between the other two, and does not have the advantage of being approximately orthogonal. Therefore, from now on, the analysis will be based on the models `SPC orthog.` and `SPC orthog. sparse`, which are complementary:

- `SPC orthog.` captures the same variance than `PCA` with just one component, but with sparsity gains. A measure of welfare consisting of **the first component** of `SPC orthog.` explains 20% of variance (a significant amount, given that we have 126 original variables) and can be constructed with **half of the original features** (63 variables can be dropped). However, to explain 30% of variability we would need to use all the original features, given that each component is not sparse enough.

- `SPC orthog.sparse` can capture greater variance while keeping the sparsity (up to 30% of variance by dropping 50 variables, or even 40% of variance by dropping 11). Therefore, this is the model to use if the main goal is to find the minimum subset K^* that explains a significant amount of the variability in welfare, for example, to design shorter surveys. Nevertheless, this comes at the expense of having to use more principal components in the analysis. For instance, **3 components** capture el 20% of variance with **35% of the original features** (85 features can be dropped). It also has interpretability advantages, as each component consists of a small subset of features (Figure 4, central panel, shows that the three first components have more than 100 variables with zero loadings each). I will use this model to find the minimum subset of EPH variables needed to explain 20% and 30% of the variability in welfare, and to understand the importance of different variable groups for welfare.

Finally, Figure 5 compares the loadings that `PCA`, `SPC orthog.` and `SPC orthog. sparse` assign to each variable in the two first principal components. The upper panels show that the direction is reversed in the first component of the sparse models. That is, the same variable that has a positive weight in `PCA`, has a negative weight in the sparse models. As this direction is arbitrary, from now on I will multiply by -1 the loadings of the variables in the first component of the sparse models to facilitate the comparison. Figure 5 shows that, in general, sparser models select those variables with greater loadings in less sparse models. However, such selection is not uniform: some selected variables (i.e. variables with non-zero loadings in sparse models) have lower `PCA` loadings than other non-selected variables (i.e. than variables with zero loadings in sparse models).

Figure 5: Features selected by sparse models -comparison of the loadings of each feature in `PCA`, `SPC orthog.`, and `SPC orthog. sparse`



5.2 The dimensions of welfare

Figure 6 summarizes PCA and SPC orthog. results in biplots. Biplots show the projection of each observation (black dots) and the original variables (colored vectors) in the two first principal components, which, in these models, capture more than 25% of the variability in welfare. If these components adequately represent the original data, two close-by points represent two people that are alike in terms of welfare. Meanwhile, vectors that point towards the same direction correspond to variables with similar response patterns across people, and that therefore have a similar meaning in terms of welfare.

Firstly, biplots show which types of variables are represented in each component and which have similar response patterns:

- Variables by topic. Different topics have different colors in the upper panel biplots. Both PCA and SPC orthog. have variables form all four groups in each component. In SPC orthog., employment variables are closer to the first component, and housing variables to the second one. There are subgroups of variables in each topic that are correlated. For instance, those in higher income deciles at the regional level, are also in higher deciles at the national level. In turn. the correlation in employment variables may be reflecting the hierarchical structure of the EPH questionnaire. There is also correlation between variables from different topics, which means that, for example, those with certain employment features usually have similar income.

- Variables by level. Variables at different levels have different colors in the central panel biplots. The first component is more related to person-level variables, while the second one, to household-level and dwelling-level variables. This means that the response profiles at the household level do not correspond to a unique response profile at the individual level: inside a household there might be individuals with different characteristics.

- Ordering in terms of welfare. Variables that constitute an ordinal welfare measure (i.e., higher values of the variable mean greater welfare levels) are purple in the inferior panel biplots. In SPC orthog., higher values in each component mean higher welfare levels. The closeness between different variables associated to greater welfare means that those people that are better off in some aspects, are also better in others.

Sparser models have a similar variable structure than PCA. Sparsity is reflected by the fact that (i) less variables are part of SPC orthog biplots, and (ii) the variables vectors are closer to the axes, so that each principal component can be more unambiguously associated with a subset of variables. In fact, when doing a biplot for SPC orthog sparse, the variable vectors lie directly on the axes, because the variables with non-zero weights in the first component have weights equal to zero in the second component and vice versa,.

Secondly, biplots show that people are grouped in two clusters. In PCA, one group is more heterogeneous, on average better positioned in terms of person-level variables, and worse off in household-level variables than the other. This can be expected if a better position at the household level compensates, for example, a lower personal income. The sparser the model, the more the clusters are differentiated along the first principal component, that is, they mainly reflect individual differences while including people from heterogeneous household backgrounds.

Figure 6: PCA and SPC orthog. biplots

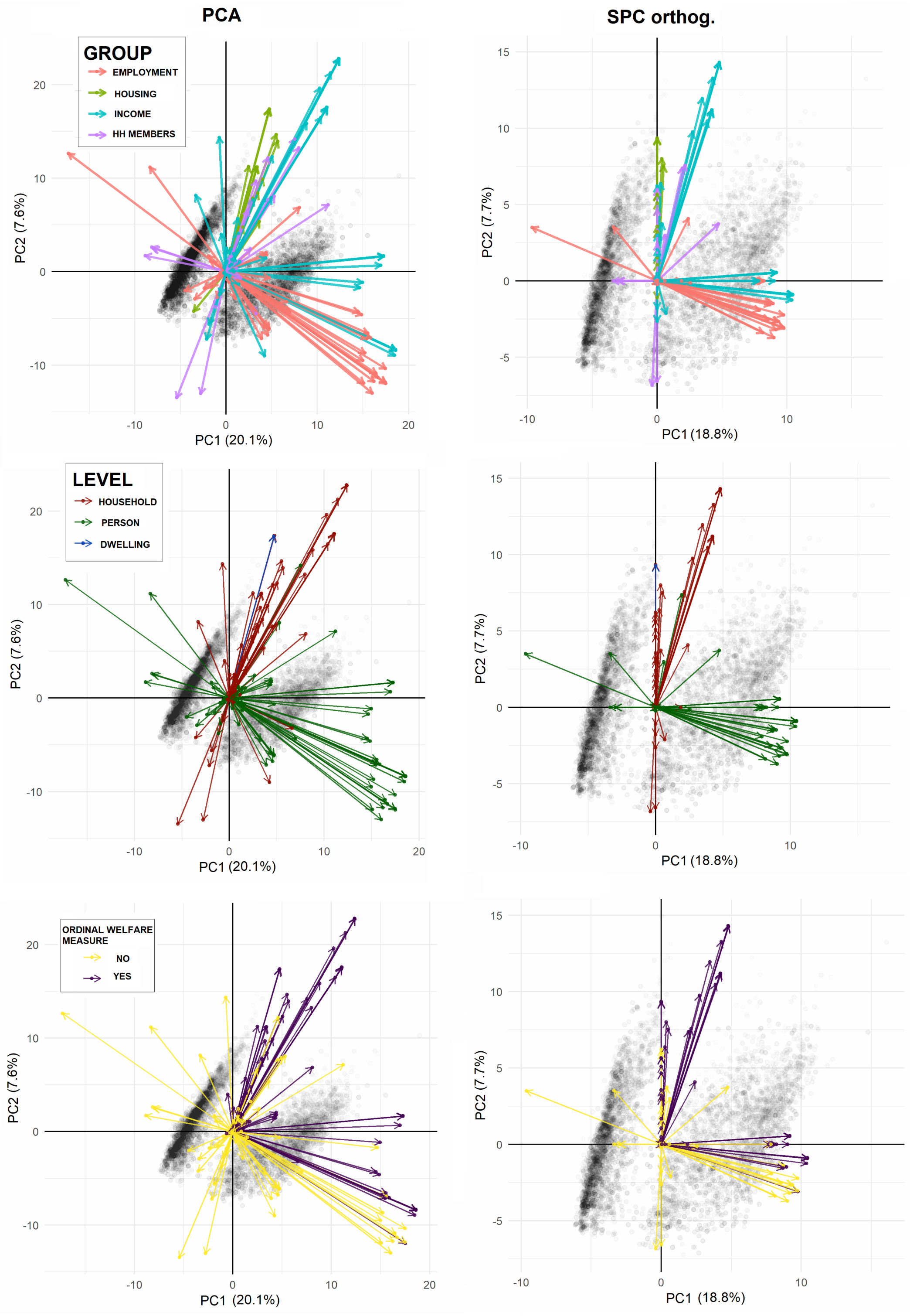
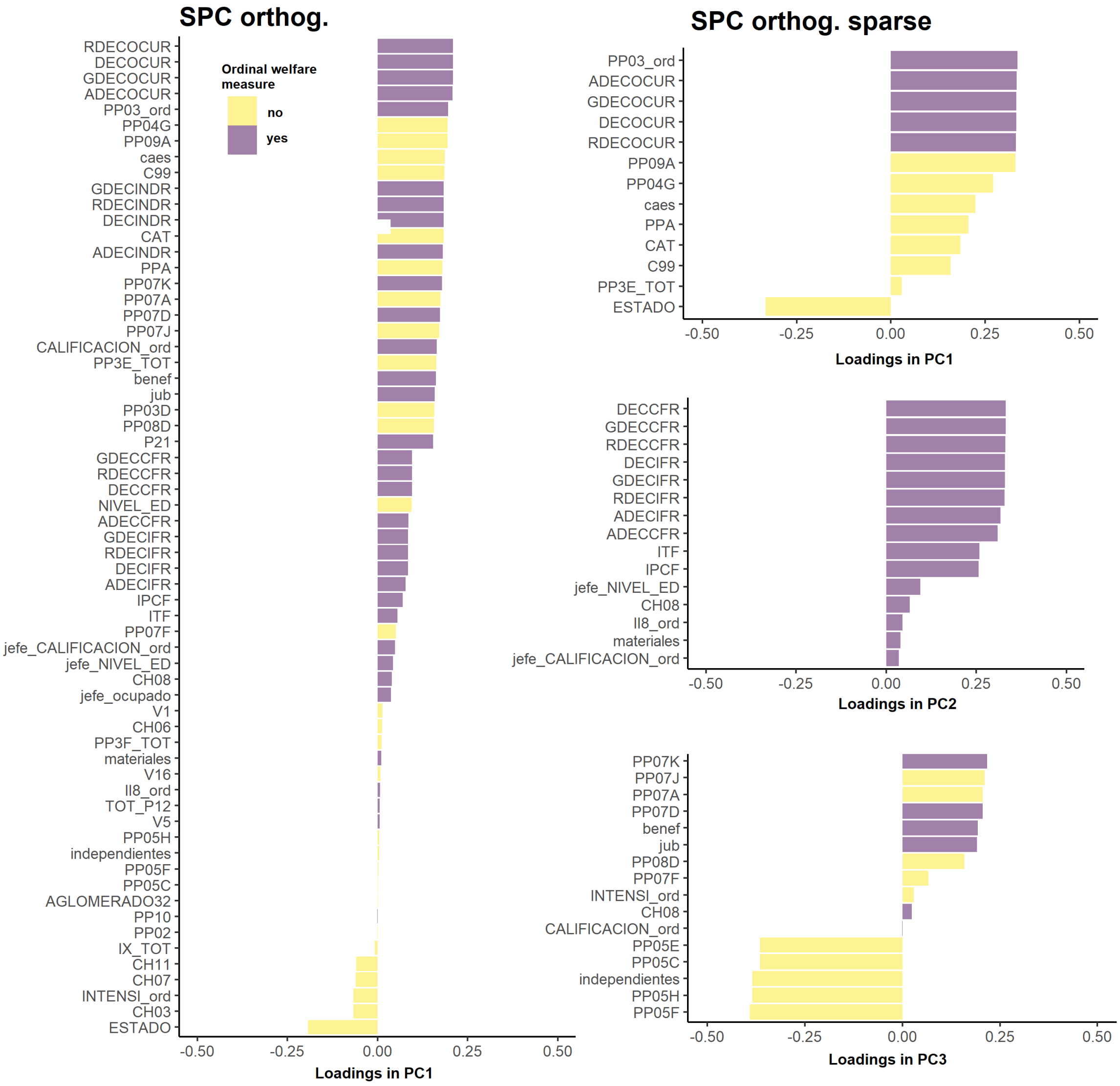


Figure 7: Ordinal versus not ordinal variables in terms of welfare, SPC orthog. and SPC orthog. sparse



5.2.1 Levels of welfare

Figure 7 plots the loadings of each variable needed to account for 20% of the original data variance: the first component of SPC orthog. and the three first components of SPC orthog. sparse. Variables that are ordinal in terms of welfare are all oriented towards the same direction of each component, which means that people that are better off in some areas, are also better in others. This is aligned with previous literature that found that different aspects of welfare are correlated (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013).

Which are the most relevant ordinal variables to explain the variability in welfare? And to which non-ordinal variables are they associated? SPC orthog. sparse, that distinguishes three main directions of variation in welfare, helps to answer these questions:

- In the first component, that captures 8% of variability, has the income deciles of the main occupation. The non-ordinal variables associated with higher income are: being employed (ESTADO), working in the City of Buenos Aires (PP09A); working in office or stores (PP04G); working in the public administration or services sector –while working in housekeeping or construction is negatively correlated with income (caes)–; being employed in the public sector (PPA); being the employer or a salaried employee (CAT); working in large companies (C99); working most of the time in the main occupation (PP3E_TOT).

- The second component, that captures 7% of variability, has ordinal variables only: total and per capita household income (deciles and levels), education of the head of the household, having medical insurance (CH08), using electricity or gas for cooking (II8_ord), having sufficient housing materials (materiales) and the qualification status of the head of the household (jefe_CALIFICACION_ord).
- Finally, variables in the third component, that captures 5% of variability, are associated to formal jobs: getting payed though formal channels (PP07K), having longer work contracts (PP07D), having benefits at work like paid vacations, bonuses, sick leave days and paid medical insurance (benef) and having pension contributions (jub).

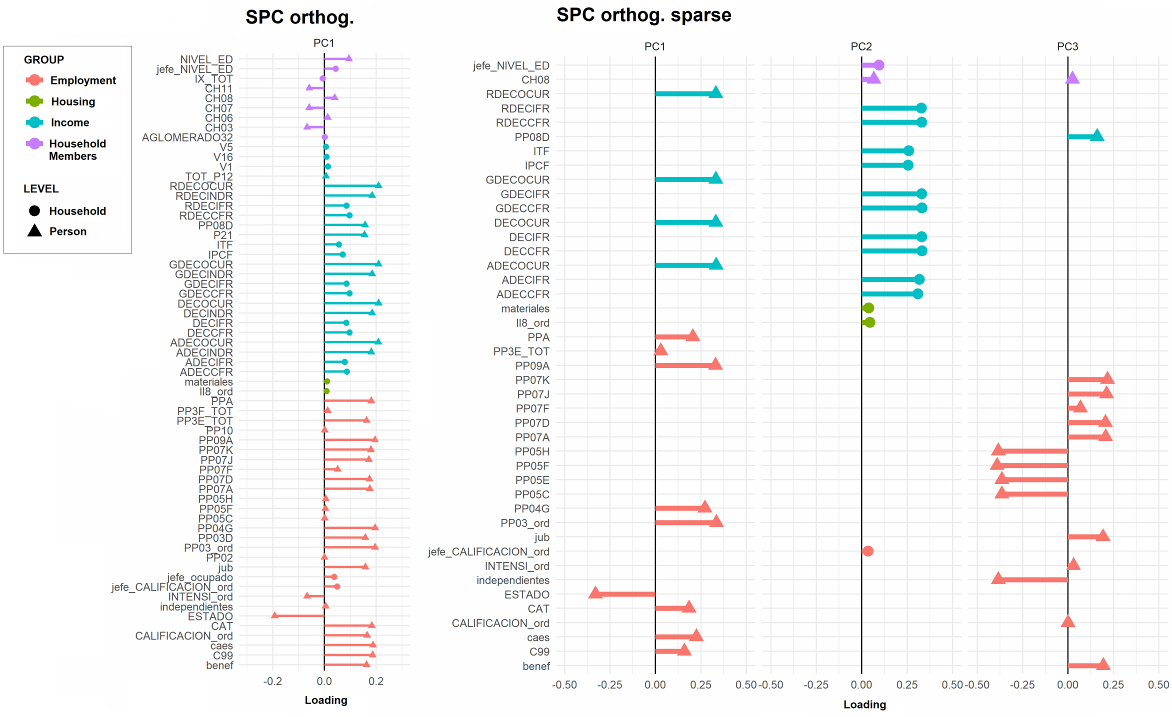
In turn, the first principal component from `SPC orthog.` might be used as a welfare index, where greater values are *generally* associated to greater welfare, but as it includes some variables that cannot be sorted according to welfare levels, it is not strictly monotonous. That is, different index levels may not mean different welfare levels. For instance, among two people in the same income decile, one who is inactive would have a lower value in the index than someone that is employed, while being inactive does not necessarily imply a lower welfare. However, as being inactive is negatively correlated with income, inactive people get lower values in the index. Therefore, the focus of this thesis is *not* to get an index of welfare levels, but rather to find groups that are different when it comes to welfare conditions, though not necessarily doing 'better' o 'worse'. In this regard, an unidimensional index using the first principal component of `SPC orthog.` or a multidimensional index using the three first components of `SPC orthog. sparse` may be useful for differentiating groups of people with median features versus extreme features when it comes to welfare, which resembles the definition of middle class by Gigliarano & Mosler (2009).

Instead, if the goal was to get an index to sort people in terms of lower or greater welfare, only orderable variables should be considered, and it would be reasonable to use the household as the unit of observation (otherwise under age people should be dropped from the sample). Next, for ensuring that greater index values correspond to greater welfare levels, there are two options: either using the absolute value of the first principal component like Edo, Sosa-Escudero & Svarc (2020) do, or constraining the loadings of SPC to be positive, which can be costly in terms of explained variance (see the results for `SPC pos. orthog.` and `SPC pos. orthog. sparse`).

5.2.2 Relevant topics

The variables needed to capture the variability in welfare, to which aspects of welfare do they refer to? Variables from the four topics –income, employment, housing, and household members– are needed to explain 20% of variability, either with the first principal component of `SPC orthog.` or the first three of `SPC orthog. sparse`. This means that welfare is multidimensional, it cannot be reduced to a single group of variables. Anyway, there is room for dimensionality reduction, and `SPC orthog. sparse` helps to find the topics that capture more variability, and more importantly, to decide which features to select within

Figure 8: Relevant variables by topic and level, `SPC orthog.` and `SPC orthog. sparse`



each topic. Most variability is captured by the income and employment variables in the first component, while the second component has mainly household member variables and housing characteristics, and the third one refers to employment questions again, distinguishing the group of independent workers from salaried workers. These are the most relevant variables withing each group:

- Income: deciles of the income from the main occupation and deciles of household income. I performed two robustness checks. Firstly, I kept just one of the variables of deciles of income (RDECOCUR), and it was still the most important. Secondly, I randomly permuted without replacement the values of RDECOCUR and it was no longer relevant.
- Employment: activity status; activity type for salaried workers (firm size, sector, whether it is located in Ciudad de Buenos Aires, establishment type, hours worked per day); formality.
- Housing: dwelling materials and energy sources.
- Household members: education of the head of the household and having health insurance.

5.2.3 Shorter surveys

This section lists the minimum subset of features $K^* \subset K$ for capturing 20% and 30% of the variability in welfare, based on `SPC orthog. sparse`. These subsets include only the 35% and the 50% of the original features, respectively. Each item lists the variables that need to be included to get each additional component, excluding those previously mentioned.

- **20% of variance** in 3 principal components, dropping 83 variables

  1. First component (8% cumulative variance): PP03_ord (would not like to work more hours), ADECOCUR, GDECOCUR, DECOCUR (deciles of income from the main occupation), ESTADO, RDECOCUR, PP09A (working in Ciudad de Buenos Aires), PP04G (working place), caes (activity subsector), PPA (activity sector), CAT (activity category), C99 (firm size), PP3E_TOT (hours worked in the main occupation).
  2. Second component (15% cumulative variance): DECCFR, GDECCFR, RDECCFR, DECIFR, GDECIFR, RDECIFR, ADECIFR, ADECCFR (deciles of household income), ITF, IPCF (household income), jefe_NIVEL_ED (education of the head of the household), CH08 (health insurance), II8_ord (energy sources), materiales (housing conditions), jefe_CALIFICACION_ord (qualifications of the head of the household).
  3. Third component (20% cumulative variance): PP05F, PP05H, independientes, PP05C, PP05E (employment features of independent workers), PP07K, PP07J, PP07A, PP07D, benef, jub, PP07F (employment features of salaried workers), PP08D (salaried income), INTENSI_ord (over-occupation and sub-occupation), CH08 (having health insurance), CALIFICACION_ord (qualifications).

- **30% of variance** in 6 principal components, dropping 64 variables. These should be added to the already mentioned:

  4. Fourth component (24% cumulative variance): hacinamiento_v, hacinamiento, IX_TOT (number of household and dwelling members), CH06 (age), W4 (percentage of household members participating in housekeeping tasks), usos (number of exclusive-use rooms over the total number of rooms used for sleeping and working), V1 (living from employment income), CH03 (relationship with the head of the household), V2 (receiving a pension), V2_M (pensions amount), IX_MEN10 (members under 10 years old in the household), CH11 (educational establishment type), W_11 (doing household chores), jefe_ocupado (whether the head of the household is employed), CH07 (marital status).
  5. Fifth component (28% cumulative variance): PP10A, PP11LO, PP10, PP02, PP11L1 , PP11PQ, PP11ST, PP11R, PP11B1 (unemployed features, PP07E11M (whether the person has worked in an internship or for an initially agreed short period of time), tiempo (months working), V3V4, V3V4_M (income from unemployment insurance), NIVEL_ED (education level).
  6. Sixth component (30% cumulative variance): ADECINDR, DECINDR, GDECINDR, RDECINDR (deciles of total personal income), IX_TOT (number of household members).

5.3 Validation

If welfare is stable over short periods of time, a way of assessing the performance of the methodology is analyzing whether its results are stable.

5.3.1 Loadings over time

To begin with, I compare the loadings for each variable in `SPC orthog.` and `SPC orthog. sparse` when estimating them with the EPH for the fourth quarter of 2019 (4T2019) versus when estimating them with

Table 2: Error metrics for the location in the first principal component of `SPC orthog.`

MSE_{4T}	Quartiles	Deciles
0.00005	$MSEQ_{4T} = 0.001$ $inacc_{4T} = 0.1\%$ $maxdist_{4T} = 1$	$MSEQ_{4T} = 0.004$ $inacc_{4T} = 0.4\%$ $maxdist_{4T} = 1$

the EPH for the third quarter of 2019 (3T2019). Figure 9 shows that, for `SPC orthog.`, they are almost exactly the same. For `SPC orthog. sparse`, the variables with non-zero loadings are generally the same, especially those with the greater absolute value of the loadings.

5.3.2 Location in the subspace of welfare

Another way of measuring stability over time is by comparing the location of each person in the new subspace of P principal components. I calculate the difference between:

- Real location $F_{4T2019} = X_{4T2019}V_{4T2019}$: location in the welfare of the fourth quarter computed with the loadings estimated with data for the *fourth* quarter.
- Predicted location $\hat{F}_{4T2019} = X_{4T2019}V_{3T2019}$: location in the welfare of the fourth quarter computed with the loadings estimated with data for the *third* quarter.

Figure 10 compares the distributions of people's locations in the first component of `PCA`, `SPC orthog.` and the three first components of `SPC orthog. sparse`. `SPC orthog.` is stabler than `PCA`. Table 2 shows the prediction error measures for `SPC orthog.` as explained in Section 3.5. The MSE is really low and smaller than the one for `PCA` (0,02). Also, the location of people in quartiles of deciles of the first principal component is quite accurate: only the 0,1% and the 0,4% of the people are misclassified, respectively, and at most they are classified in the adjacent quantile.

In turn, Table 3 shows the prediction error metrics for the locations in the space of the three principal components of `SPC orthog. sparse`. To calculate depth I dropped duplicated observations. The left and central panels of Figure 11 plot the locations of a random subsample of 1000 people using a different color for each depth quartile. The spatial data configuration is not suitable for defining a unique centrality measure and calculating distance from it, because there are clusters. Therefore, quartiles based on Tukey's depth are quite imprecise, 14,9% of people are misclassified, many to far away quartiles. In turn, quartiles based on the Mahalanobis' distance capture one of the clusters, and improve the classification, with just 3,5% of people misclassified and to adjacent quartiles at most. Nevertheless, only 0,7% of people is misclassified under *both* depth measures, which means that the imprecision comes from the quartile definition rather than from variations of data over time. To solve this issue, I grouped individuals using k-means, which is able to capture the spatial configuration of data, as the right panel in Figure 11 shows. The latter is really stable over time, with only 0,1% misclassified (Table 3 and Table 4).

In summary, results are stable over time, both for the location in the first principal component of `SPC orthog.`, and for the location in the space of three principal components of `SPC orthog. sparse`.

Figure 9: Stability of loadings over time

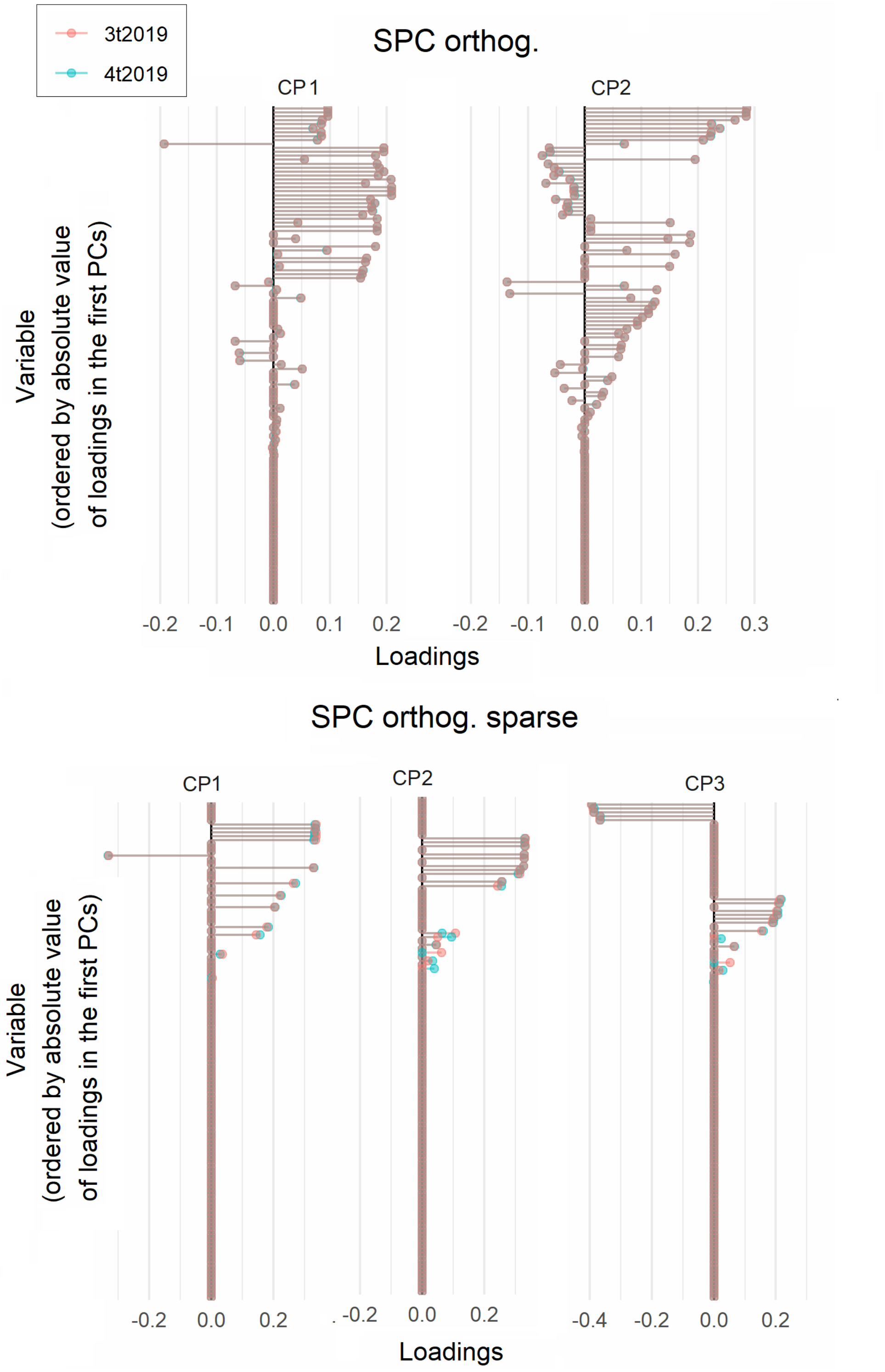


Figure 10: Stability of the distributions of locations of people in the principal components over time

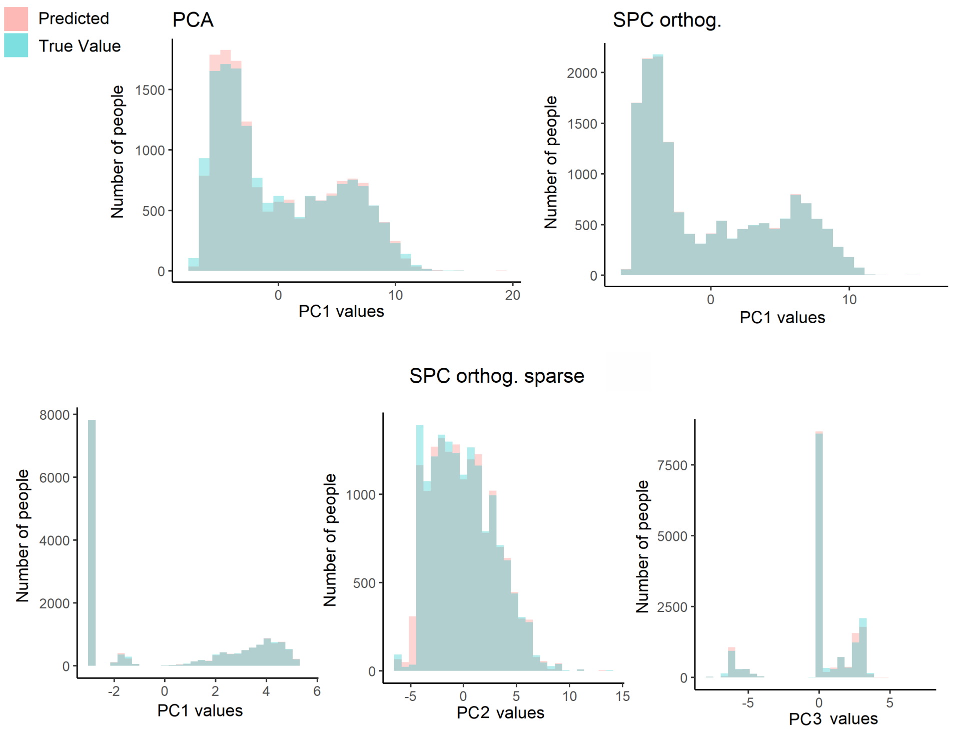


Figure 11: Location in the space of the three principal components of `SPC orthog. sparse`

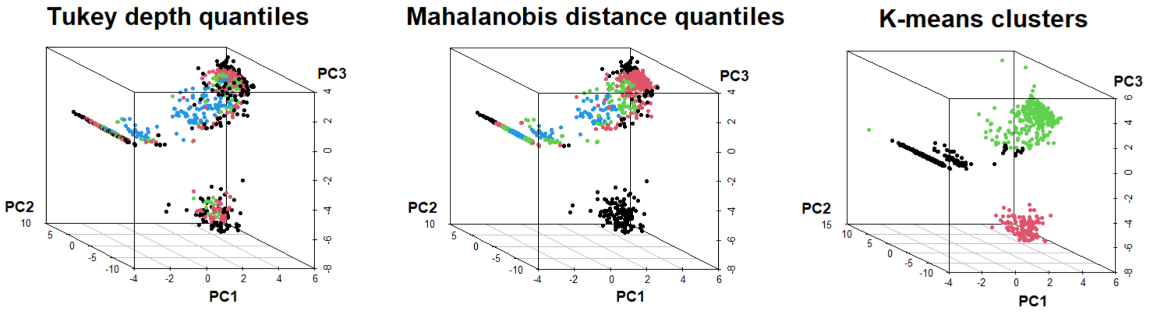


Table 3: Error metrics for the location in the three principal components of `SPC orthog. sparse`

MSE_{4T}	Quartiles Tukey's depth	Quartiles Mahalanobis distance	K-means clusters
0.0042	$MSEQ_{4T} = 0.16$ $inacc_{4T} = 14,9\%$ $maxdist_{4T} = 3$	$MSEQ_{4T} = 0.035$ $inacc_{4T} = 3.5\%$ $maxdist_{4T} = 1$	$inacc_{4T} = 0.1\%$

Table 4: Classification in k-means clusters

		Predicted cluster		
		1	**2**	**3**
Real cluster	**1**	8882	0	13
	2	0	1843	0
	3	4	0	4868

6 Conclusions

Selecting the smallest set of variables needed for measuring welfare would allow to design shorter, faster to implement, and less expensive surveys, and to lower the non-response rates. This thesis implements PCA with sparse loadings to find the dimensionality of welfare, as captured by the Argentinian EPH, and to select the smallest set of variables for measuring it. This methodology has two advantages over PCA in this specific task. The first is that it explains a similar amount of the variability in welfare using less features. The second is its interpretability: each dimension of welfare can be associated to a subset of the original variables. Therefore, PCA with sparse loadings is useful for exploring the multidimensionality of welfare hypothesis, that is, for finding how many interpretable dimensions are needed for capturing it. At the same time, optimal scaling or non-linear PCA techniques make it possible to use 126 variables from the EPH as a starting point, avoiding ad-hoc selections, despite most of those variables are of mixed types, mainly categorical.

Although welfare is indeed multidimensional, there is room for dimensionality reduction. On the one hand, the trade-off between the explained variance and the sparsity of the components is a sign of the multidimensionality. The greater this trade-off, the more variables should be used for explaining a good portion of welfare, which means that welfare is 'more' multidimensional. The results are aligned with previous literature: income variables matter, but also do others, mainly the employment ones. To explain 20% of the variability in welfare, many aspects are important: income, employment, housing characteristics, education, and health. On the other hand, it is possible to explain that 20% of variance with just 35% of the original features, or to explain 30% using just half of them, which allows to design shorter surveys. Among each topic, the selected variables are: main occupation income deciles and household income deciles; main activity and type of activity of salaried workers; construction materials and energy sources among household characteristics; educational level of the head of the household and having health insurance. Most variability in welfare is related to the level of income in the main occupation, which in turn correlates with being employed in certain activities (for example, geographic localization, activity sector, firm size).

Also in accordance with previous findings, the variables that can be ordered in terms of welfare are correlated: people with higher welfare in some aspects also are better off in others. Additionally, person-level variables are captured by the first principal component and household variables at the person level in the second one, which means that there is heterogeneity within the same household. Finally, the methodology is stable over time.

These results are relative to surveys similar to the EPH, which does not necessary cover all relevant welfare aspects, such as subjective features (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013). Deciding on the best data source to use as a starting point is out of the scope of this thesis and can be explored in future work. Other future research lines are, first, to create indexes that identify groups of different welfare levels. The relevant features to capture variability in general welfare do not need to be the same as the ones needed to measure poverty or wealth (Edo, Sosa-Escudero & Svarc, 2020). Supervised methodologies to select principal components, such as Lassoed PCA by Witten & Tibshirani (2008) might be useful for that aim. Secondly, the fact that person-level and household-level features are captured in different principal components might be useful for studying social mobility overtime by looking at movements in each of the directions. Finally, many of the variables selected by the sparse models are correlated, so there is room for working in a method to select just some of them. An option might be to find principal components in each group of correlated variables as a previous step, which may also be useful for better summarizing the hierarchical structure of the EPH.

7 Bibliography

Aaberge, R., & Brandolini, A. (2015). Multidimensional poverty and inequality. In *Handbook of Income Dstribution* (Vol. 2, pp. 141-216). Elsevier.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073-1076.

Cadima, J., & Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2), 203-214.

Caruso, G., Sosa-Escudero, W., & Svarc, M. (2015). Deprivation and the dimensionality of welfare: a variable-selection cluster-analysis approach. *Review of Income and Wealth*, 61(4), 702-722.

Coromaldi, M., & Zoli, M. (2012). Deriving multidimensional poverty indicators: Methodological issues and an empirical analysis for Italy. *Social indicators research*, 107(1), 37-54.

Edo, M., Sosa-Escudero, W., & Svarc, M. (2021). A multidimensional approach to measuring the middle class. *The Journal of Economic Inequality*, 19(1), 139-162.

Ferro Luzzi, G., Flückiger, Y., & Weber, S. (2008). A cluster analysis of multidimensional poverty in Switzerland. In *Quantitative approaches to multidimensional poverty measurement* (pp. 63-79). Palgrave Macmillan, London.

Fraiman, R., Justel, A., & Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, 103(483), 1294-1303.

Gasparini, L., Sosa-Escudero, W., Marchionni, M., & Olivieri, S. (2013). Multidimensional poverty in Latin America and the Caribbean: new evidence from the Gallup World Poll. *The Journal of Economic Inequality*, 11(2), 195-214.

Gigliarano, C., & Mosler, K. C. (2009). Measuring middle-class decline in one and many attributes. *Università Politecnica delle Marche, Dipartimento di economia.*

Gimenez, Y. (2015). *Selección de variables para datos multivariados and datos funcionales.* Tesis doctoral. Facultad de Ciencias Exactas and Naturales. Universidad de Buenos Aires.

Gómez, A., Álvarez, G., Mario, S., & Olmos, F. (2004). Metodología de elaboración del Índice de Privación Material de los Hogares (IPMH). *Serie Pobreza. INDEC. DNESyP/DEP/P5/PID*

INDEC (2019). Indicadores de condiciones de vida de los hogares en 31 aglomerados urbanos. Primer semestre de 2019. *Informes Técnicos* 3(204). ISSN 2545-6636. *Condiciones de vida* 3(15). ISSN 2545-6660.

INDEC (2020). Encuesta Permanente de Hogares. Diseño de registro and estructura para las bases preliminares Hogar and Personas.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790-794.

Kakwani, N., & Silber, J. (Eds.). (2008). *Quantitative approaches to multidimensional poverty measurement.* Springer.

Kozlowski, D., Tiscornia, P., Weksler, G., Rosati, G. & Shokida, N. (2020). eph: Argentina's Permanent Household Survey Data and Manipulation Utilities. *R package version.* https://doi.org/10.5281/zenodo.3462677

Mair, P., & de Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: The R package aspect. *Journal of Statistical Software*, 32(1), 1-23.

Mair, P., & de Leeuw, J. (2018). aspect: A General Framework for Multivariate Analysis with Optimal Scaling. *R package version 1.0-5.* https://CRAN.R-project.org/package=aspect

Merola, G. M., & Baulch, B. (2019). Using sparse categorical principal components to estimate asset indices: new methods with an application to rural Southeast Asia. *Review of Development Economics*, 23(2), 640-662.

Mori, Y., Kuroda, M., & Makino, N. (2016). *Nonlinear principal component analysis and its applications.* New York: Springer.

Sen, A. (1985). *Commodities and Capabilities.* Oxford: Oxford University Press.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3), 515-534.

Witten, D. M., & Tibshirani, R. (2008). Testing significance of features by lassoed principal components. *The annals of applied statistics*, 2(3), 986.

Witten, D. M., & Tibshirani, R. (2020). PMA: Penalized Multivariate Analysis. *R package version 1.2.1.* https://CRAN.R-project.org/package=PMA

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265-286.

Zou, H., & Hastie, T., (2020). elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. *R package version 1.3.* https://CRAN.R-project.org/package=elasticnet

A Annex: list of variables

Name	Group	Level	Welfare	Description
ADECCFR	Income	Household	yes	Ídem INDEC (2020)
ADECIFR	Income	Household	yes	Ídem INDEC (2020)
ADECINDR	Income	Person	yes	Ídem INDEC (2020)
ADECOCUR	Income	Person	yes	Ídem INDEC (2020)
AGLOMERADO32	Household Members	Household	no	Lives in Ciudad Autónoma de Buenos Aires
agua	Housing	Household	yes	Access to water (INDEC, 2019), used by Edo, Sosa-Escudero & Svarc (2020)
ayuda	Income	Household	no	Receives in-kind assistance from people or institutions outside the household
benef	Employment	Person	yes	Has vacation days, bi-annual bonuses, sick days, health insurance (how many checked)
C99	Employment	Person	no	Number of people at current or previous job
caes	Employment	Person	no	Industry Classification
CALIFICACION_ord	Employment	Person	yes	Employment classification
CAT	Employment	Person	no	Groups CAT_OCUP and CAT_INAC from INDEC (2020)
CH03	Household Members	Person	no	Ídem INDEC (2020)
CH04	Household Members	Person	no	Ídem INDEC (2020)
CH06	Household Members	Person	no	Ídem INDEC (2020)
CH07	Household Members	Person	no	Ídem INDEC (2020)
CH08	Household Members	Person	yes	Ídem INDEC (2020)
CH09	Household Members	Person	yes	Ídem INDEC (2020)
CH10_ord	Household Members	Person	yes	Attends an educational establishment (if younger than 17); attends or attended (17 years old or older)
CH11	Household Members	Person	no	Ídem INDEC (2020)
CH15	Household Members	Person	no	Ídem INDEC (2020)
CH16	Household Members	Person	no	Ídem INDEC (2020)
DECCFR	Income	Household	yes	Ídem INDEC (2020)
DECIFR	Income	Household	yes	Ídem INDEC (2020)
DECINDR	Income	Person	yes	Ídem INDEC (2020)
DECOCUR	Income	Person	yes	Ídem INDEC (2020)
ESTADO	Employment	Person	no	Ídem INDEC (2020)
GDECCFR	Income	Household	yes	Ídem INDEC (2020)
GDECIFR	Income	Household	yes	Ídem INDEC (2020)
GDECINDR	Income	Person	yes	Ídem INDEC (2020)
GDECOCUR	Income	Person	yes	Ídem INDEC (2020)
hacinamiento	Housing	Household	yes	Number of rooms for exclusive use of the household over total household members
hacinamiento_v	Housing	Dwelling	yes	Number of rooms for exclusive use in the dwelling over total household members
II4	Housing	Household	yes	Number of rooms that are kitchen, laundry room, or garage
II7	Housing	Household	no	Ídem INDEC (2020)
II7_propietario	Income	Household	yes	Owns the dwelling and/or land, used by Edo, Sosa-Escudero & Svarc (2020)
II8_ord	Housing	Household	yes	Fuel used for cooking: electricity or mains gas; bottled gas; kerosene, wood or charcoal
II9_ord	Housing	Household	yes	Ídem INDEC (2020)

Name	**Group**	**Level**	**Welfare**	**Description**
independientes	Employment	Person	no	Type of company: without partners, legal partnership, other family partnership, other non-family partnership.
INTENSI_ord	Employment	Person	no	(Not) over- or under-employed due to insufficient working hours
IPCF	Income	Household	yes	Ídem INDEC (2020)
ITF	Income	Household	yes	Ídem INDEC (2020)
IV1	Housing	Household	no	Dwelling type
IV1_ord	Housing	Household	yes	Lives in a house or department
IV12	Housing	Household	yes	Location in risky area, used by Edo, Sosa-Escudero & Svarc (2020)
IX_MEN10	Household Members	Household	no	Ídem INDEC (2020)
IX_TOT	Household Members	Household	no	Ídem INDEC (2020)
jefe_CALIFICACION_ord	Employment	Household	yes	Qualified head of household, used by Edo, Sosa-Escudero & Svarc (2020)
jefe_CH09	Household Members	Household	yes	Literate head of household, used by Edo, Sosa-Escudero & Svarc (2020)
jefe_NIVEL_ED	Household Members	Household	yes	Educational level of the head of the household, used by Edo, Sosa-Escudero & Svarc (2020)
jefe_ocupado	Employment	Household	yes	Employed head of the household, used by Edo, Sosa-Escudero & Svarc (2020)
jub	Employment	Person	yes	Pension contribution at current or last job, private contribution, or no contribution
materiales	Housing	Household	yes	Index of material deprivation of households (INDEC, 2019), used by Edo, Sosa-Escudero & Svarc (2020)
NIVEL_ED	Household Members	Person	no	Ídem INDEC (2020)
P21	Income	Person	yes	Ídem INDEC (2020)
PP02	Employment	Person	no	Groups PP02C and PP02E from INDEC (2020)
PP02H	Employment	Person	no	Ídem INDEC (2020)
PP02I	Employment	Person	no	Ídem INDEC (2020)
PP03_ord	Employment	Person	yes	Groups PP03G, PP03GH, PP03GI and PP03GJ from INDEC (2020): the person did not wanted to work for more hours (during the week, month, find another job), wanted and could wanted but could not, etc.
PP03D	Employment	Person	no	Ídem INDEC (2020)
PP04B2	Employment	Person	no	Ídem INDEC (2020)
PP04G	Employment	Person	no	Ídem INDEC (2020)
PP05C	Employment	Person	no	Owns machinery, establishment or vehicle
PP05E	Employment	Person	no	Ídem INDEC (2020)
PP05F	Employment	Person	no	Ídem INDEC (2020)
PP05H	Employment	Person	no	Ídem INDEC (2020)
PP07A	Employment	Person	no	Ídem INDEC (2020)
PP07D	Employment	Person	yes	Ídem INDEC (2020)
PP07E11M	Employment	Person	no	Groups PP07E and PP11M from INDEC (2020)
PP07F	Employment	Person	no	Receives free food, housing, any product or service, another benefit (how many checked)
PP07J	Employment	Person	no	Ídem INDEC (2020)
PP07K	Employment	Person	yes	Ídem INDEC (2020)
PP08D	Income	Person	no	Groups PP08D1 and PP0D4 de INDEC (2020)
PP08F	Income	Person	no	Groups PP08F1 and PP08F2 de INDEC (2020)
PP08J1	Income	Person	yes	Ídem INDEC (2020)

Name	Group	Level	Welfare	Description
PP08J2	Income	Person	no	Ídem INDEC (2020)
PP08J3	Income	Person	no	Ídem INDEC (2020)
PP09A	Employment	Person	no	Ídem INDEC (2020)
PP10	Employment	Person	yes	Ídem INDEC (2020)
PP10A	Employment	Person	yes	Ídem INDEC (2020)
PP11B1	Employment	Person	no	Ídem INDEC (2020)
PP11L1	Employment	Person	no	Ídem INDEC (2020)
PP11LO	Employment	Person	no	Groups PP11L and PP11O de INDEC (2020)
PP11PQ	Employment	Person	no	Groups PP11P and PP11Q de INDEC (2020)
PP11R	Employment	Person	no	Ídem INDEC (2020)
PP11ST	Employment	Person	yes	Groups PP11S and PP11T de INDEC (2020)
PP3E_TOT	Employment	Person	no	Ídem INDEC (2020)
PP3F_TOT	Employment	Person	no	Ídem INDEC (2020)
PPA	Employment	Person	no	Groups PP04A and PP11A de INDEC (2020)
RDECCFR	Income	Household	yes	Ídem INDEC (2020)
RDECIFR	Income	Household	yes	Ídem INDEC (2020)
RDECINDR	Income	Person	yes	Ídem INDEC (2020)
RDECOCUR	Income	Person	yes	Ídem INDEC (2020)
saneamiento	Housing	Household	yes	Access to sanitation (INDEC, 2019), used by Edo, Sosa-Escudero & Svarc (2020)
tiempo	Employment	Person	no	How long has the person worked there
TOT_P12	Income	Person	yes	Ídem INDEC (2020)
usos	Housing	Household	yes	Total rooms with an exclusive use over those used for sleeping and working
usos_II4	Housing	Household	yes	Number of rooms that are kitchen, laundry room or garage and not used for sleeping
V1	Income	Household	no	Lived on work earnings
V10	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V10_M	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V11_M	Income	Household	no	Ídem INDEC (2020)
V12_M	Income	Household	no	Ídem INDEC (2020)
V13	Income	Household	no	Ídem INDEC (2020)
V14	Income	Household	no	Ídem INDEC (2020)
V15	Income	Household	no	Ídem INDEC (2020)
V16	Income	Household	no	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V17_ord	Income	Household	yes	Ídem INDEC (2020)
V18_M	Income	Household	no	Ídem INDEC (2020)
V2	Income	Household	no	Lived on retirement funds or pension received this or last month
V2_M	Income	Household	no	Ídem INDEC (2020)
V21_M	Income	Household	yes	Ídem INDEC (2020)
V3V4	Income	Household	no	Lived on severance pay or unemployement insurance
V3V4_M	Income	Household	no	Amount received for severance pay and unemployemnt ensurance
V5	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V5_M	Income	Household	no	Ídem INDEC (2020)
V8	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)

Name	**Group**	**Level**	**Welfare**	**Description**
V8_M	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V9	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
V9_M	Income	Household	yes	Ídem INDEC (2020), used by Edo, Sosa-Escudero & Svarc (2020)
W_11	Household Members	Person	no	Does the household chores
W_21	Household Members	Person	no	Helps with household chores
W_domestico	Household Members	Household	yes	The household has a maid, used by Edo, Sosa-Escudero & Svarc (2020)
W_externo	Household Members	Household	no	Someone from outside the household does or helps with household chores
W4	Household Members	Household	no	Number of people doing household chores over total household members
W5	Household Members	Household	no	Number of people heloping household chores over total household members