Universidad de
SanAndrés

Universidad de San Andrés
Departamento de Economía
Maestría en Economía

¿Cuán multidimensional es el bienestar? Un análisis de componentes principales esparsos

Wendy BRAU
DNI 39.242.811

Mentor: Walter Sosa-Escudero

Ciudad Autónoma de Buenos Aires
10 de diciembre, 2021

Tesis de Maestría en Economía de
Wendy BRAU

"¿Cuán multidimensional es el bienestar? Un análisis de componentes principales esparsos"

Resumen

Este trabajo trata de medir la dimensionalidad del bienestar aplicando técnicas de Análisis de Componentes Principales (PCA) con pesos esparsos, que combinan PCA con técnicas de regularización, y usa PCA no lineal para trabajar con datos mixtos. Asumiendo que el bienestar puede representarse con un subespacio de determinado conjunto de datos, la hipótesis de multidimensionalidad del bienestar refiere a que más de una dimensión interpretable es necesaria para caracterizarlo. Una aplicación empírica a la Encuesta Permanente de Hogares expone las limitaciones de PCA y las ventajas de usar PCA con pesos esparsos a la hora de determinar el subconjunto relevante de variables para medir el bienestar. Partiendo de 126 variables numéricas y categóricas, determino dicho subconjunto, lo que permitiría implementar encuestas más breves. Encuentro que el bienestar es multidimensional, pero que hay espacio para reducir la dimensión: con tres componentes principales esparsos, es posible explicar un 20 % de variabilidad usando sólo el 35 % de las variables originales, y el 30 % de la variabilidad usando la mitad. Con un solo componente principal esparso, es posible capturar el 20 % de la variabilidad en el bienestar usando la mitad de las variables.

Palabras clave: Bienestar, Análisis de Componentes Principales, PCA esparso, PCA no lineal, Regularización, Encuesta Permanente de Hogares, Argentina.

"How multidimensional is welfare? A sparse principal components analysis"

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 subset among 126 mixed type variables, which could be useful for implementing shorter surveys. 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.

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

Códigos JEL: C38, I31, C55

Índice

1. Introducción

Aunque hay acuerdo en la literatura acerca del carácter multidimensional del bienestar (Sen, 1985; Kakwani & Silber, 2008; Aaberge & Brandolini, 2015), en la práctica es necesario saber cuántas y cuáles variables deben ser consideradas para caracterizarlo. Identificar el mínimo conjunto de variables asociadas al bienestar, además, permitiría diseñar encuestas más cortas, más rápidas de implementar, menos costosas y con menos tasas de no respuesta (Edo, Sosa-Escudero & Svarc, 2020). Asumiendo que el bienestar puede representarse a partir de cierto conjunto de datos con K variables, la hipótesis de multidimensionalidad conlleva dos preguntas, relacionadas entre sí, pero distintas. La primera es si el bienestar puede resumirse en una única dimensión o cuántas dimensiones son necesarias. De manera más formal, esta pregunta refiere a si es posible proyectar el conjunto de datos inicial en un espacio de dimensión $P < K$ que capture adecuadamente su variabilidad. La segunda, si las dimensiones que permitan captarlo tienen algún significado o interpretación conducente. En otras palabras, si cada dimensión puede asociarse a un aspecto determinado del bienestar, es decir, a un subconjunto de las variables originales (por ejemplo, variables de ingreso, o de empleo). En conjunto, la dimensionalidad del bienestar refiere a cuántas dimensiones interpretables se necesitan para caracterizarlo. Diríamos que el bienestar es unidimensional si pudiésemos proyectarlo en un espacio de una dimensión que capture adecuadamente su variabilidad y esa dimensión estuviera compuesta por una única variable o un único grupo de variables, por ejemplo, el ingreso.

El Análisis de Componentes Principales (PCA de aquí en más por sus siglas en inglés, *Principal Component Analysis*) es útil para responder a la primera pregunta. PCA es una técnica de reducción de dimensión que busca resumir la variabilidad de un conjunto de datos de K variables en $P < K$ componentes principales. Los componentes principales son las direcciones de máxima variabilidad en el conjunto de datos, en orden decreciente y ortogonales entre sí, generados a partir de una combinación lineal de las variables originales. Un subconjunto de P componentes que explique buena parte de la variabilidad en los datos originales conforma un espacio de menor dimensión que los representa adecuadamente. Por lo tanto, proyectar linealmente un conjunto de datos que captura el bienestar usando PCA puede servir para resumirlo en menos dimensiones. La linealidad de la proyección y la ortogonalidad de los componentes en PCA tienen ventajas en términos de interpretación. Por un lado, si cada componente es una combinación lineal de las variables originales, podemos recuperar los pesos de cada variable en cada componente para ver a qué subconjunto de las variables originales está asociado. Por el otro, la ortogonalidad asegura que los componentes capturen aspectos del bienestar que no están correlacionados entre sí.

Sin embargo, PCA tiene una desventaja: por lo general, cada componente principal es una combinación lineal de todas las variables originales. Primero, esto trae problemas en términos de interpretación, dado que no permite asociar cada componente a un aspecto determinado del bienestar. Segundo, si todas las variables deben usarse para reconstruir los componentes, no es posible determinar un subconjunto de pocas variables que capturen el bienestar que pueda usarse para diseñar encuestas más cortas. En cambio, los métodos de PCA con pesos esparsos, como los propuestos por Zou, Hastie & Tibshirani (2006) y Witten, Tibshirani & Hastie (2009), incorporan técnicas de selección de variables en PCA y construyen cada componente principal como combinación lineal de solamente un subconjunto de las variables originales. Esto es, construyen componentes con pesos esparsos. Por lo tanto, estos métodos son útiles para responder a la segunda pregunta sobre la dimensionalidad del bienestar, es decir, para encontrar componentes interpretables, asociados a solo un subconjunto de las variables originales. Además, tomando solamente alguno(s) de los componentes con pesos esparsos para explicar la variabilidad en los datos, es posible reducir la dimensión del espacio

original de variables y encontrar un subconjunto relevante de variables originales $K^* \subset K$ que capture la variabilidad en el bienestar.

Asumiendo que el bienestar puede representarse con un subespacio de determinado conjunto de datos de K variables, el objetivo de este trabajo es identificar el subconjunto $K^* \subset K$ relevante para explicar la variabilidad en el bienestar. Para hacerlo, se explora una metodología no usada hasta el momento con tal objetivo: PCA con pesos esparsos (Zou, Hastie & Tibshirani, 2006; Witten, Tibshirani & Hastie, 2009). Una aplicación empírica a los datos de la Encuesta Permanente de Hogares (EPH de aquí en adelante) expone las ventajas de esta metodología. Aunque la EPH refleja algunos aspectos monetarios y no monetarios del bienestar, hay aspectos subjetivos relevantes que no están reflejados (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013). Es decir, la EPH no cubre todas las dimensiones del bienestar. Determinar el conjunto de datos relevante del cual partir queda por fuera del alcance de este trabajo. Por lo tanto, los resultados serán relativos a encuestas de este tipo, serán útiles para encontrar las variables relevantes para hacerlas más breves, y referirán a la dimensionalidad de los aspectos monetarios y no monetarios del bienestar que la EPH sí captura.

La EPH tiene variables de tipo mixto: aunque algunas son numéricas, la mayoría son categóricas, tanto nominales como ordinales. Como PCA y PCA con pesos esparsos son técnicas pensadas para variables numéricas, otra de las contribuciones de este trabajo es explorar el uso de técnicas de PCA no lineal, o PCA con escalamiento óptimo (Mori, Koruda & Makino, 2016) para cuantificar los datos categóricos (i.e., asignar un número a cada una de las categorías). Esto permite trabajar con muchas más variables de la EPH y evitar una selección previa ad-hoc. En particular, usando los microdatos de las EPH correspondientes al tercer y cuarto trimestre de 2019 para la región del Gran Buenos Aires, construí dos bases a nivel individual, pero que también incluyen características del hogar de cada individuo. Cada una tiene un total 126 variables, que fueron óptimamente escaladas. Luego, implementé PCA y distintos modelos de PCA con pesos esparsos. Los distintos modelos de PCA con pesos esparsos se diferencian en (i) la metodología usada para construirlos: implemento *Sparse Principal Component Analysis* (SPCA) propuesto por Zou, Hastie & Tibshirani (2006) y dos variantes de *Sparse Principal Components* (SPC) propuestas por Witten, Tibshirani & Hastie (2009); (ii) los hiperparámetros de esparsitud usados. Finalmente, validé la metodología analizando la estabilidad en el tiempo de los resultados.

En lo que sigue, la siguiente sección discute con más detalle las diferencias y ventajas de los métodos propuestos a la luz de los trabajos previos. La tercera sección describe los métodos usados y la estrategia de validación. La cuarta sección describe los datos utilizados y su preprocesamiento, incluyendo las modificaciones producidas por el escalamiento óptimo. La quinta sección presenta tres tipos de resultados: primero, presenta los criterios de elección de modelos entre los distintos estimados; segundo, el análisis de la dimensionalidad del bienestar en base los modelos elegidos; tercero, la validación de la metodología. Las conclusiones, al final.

2. Literatura previa

Este trabajo se inscribe en las contribuciones del aprendizaje automático y estadístico a los estudios de la pobreza, la desigualdad y el desarrollo. En particular, en las contribuciones a la literatura sobre la multidimensionalidad del bienestar. Hay acuerdo acerca del carácter multidimensional del bienestar (Sen, 1985; Kakwani & Silber, 2008; Aaberge & Brandolini, 2015), esto es, acerca de la incapacidad de medirlo

usando una única dimensión interpretable. Ahora bien, en la práctica, es necesario definir cuántas de las muchas posibles variables son efectivamente relevantes y deben ser consideradas para una caracterización acertada del bienestar (Caruso, Sosa-Escudero & Svarc, 2015). A falta de esta definición, se suele utilizar el ingreso para definir grupos de bienestar, lo que podría estar dejando de lado otros aspectos relevantes. Identificar el mínimo conjunto de variables necesarias para caracterizar al bienestar, además, permitiría diseñar encuestas más cortas, más rápidas de implementar, menos costosas y con menos tasas de no respuesta.

Técnicas de aprendizaje no supervisado (es decir, sin una variable dependiente asociada) como análisis de factores y clustering han sido usadas en trabajos previos para tratar de definir la dimensionalidad del bienestar. La mayoría de ellos concluye que, aunque el ingreso suele ser una variable relevante, el bienestar es efectivamente multidimensional y otras variables también deben ser consideradas. Gasparini, Sosa-Escudero, Marchionni & Olivieri (2013) aplican análisis de factores a 12 variables asociadas al bienestar de la *Gallup World Poll* para Latinoamérica. Es decir, buscan representarlo en factores latentes que resuman las múltiples variables iniciales en un nuevo espacio de menor dimensión, e interpretan la multidimensionalidad como la necesidad de retener más de un factor. Concluyen que deben retener al menos tres factores, el primero de los cuales está asociado principalmente con el ingreso, el segundo asociado al bienestar subjetivo y el tercero a necesidades básicas. Otro trabajo que aplica análisis de factores para determinar las dimensionalidad del bienestar es Ferro Luzzi, Flückiger, & Weber (2008) a partir de 32 variables del Panel de Hogares Suizos, y concluyen que cuatro factores latentes (que reflejan aspectos financieros, de salud, del barrio y de exclusión social) son necesarios. Luego, aplican análisis de clusters sobre los factores para identificar a la población pobre.

Caruso, Sosa-Escudero & Svarc (2015) también se enfocan en el carácter multidimensional de la pobreza. Buscan determinar un subconjunto estricto de las variables originales que permita identificar al conjunto de individuos pobres. Para hacerlo, proponen usar una metodología en dos pasos. Primero, definen clusters de pobres y no-pobres usando el algoritmo de k-medias a partir de 15 variables de la *Gallup World Poll* para Latinoamérica (la mayoría son las que usan Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013). Segundo, reducen la dimensión identificando el menor conjunto de variables que puede reproducir con precisión dicha clasificación entre pobres y no-pobres. Para hacerlo, usan métodos de selección de variables en análisis de clusters de Fraiman, Justel & Svarc (2008). Esto les permite interpretar fácilmente los resultados, ya que identifican un subconjunto estricto de las variables originales. Encuentran que las variables necesarias para identificar al grupo de pobres son tres: el ingreso del hogar, tener suficiente dinero para comprar comida, y tener una computadora en el hogar. Nuevamente, el ingreso es una variable relevante, pero no la única.

Edo, Sosa-Escudero & Svarc (2021) buscan determinar las dimensiones relevantes para identificar a la clase media, y distinguirla del grupo de ricos y del grupo de pobres. Parten de 19 variables a nivel de hogar obtenidas a partir de la EPH entre 2004 y 2014, ordenables en términos de bienestar y referidas a distintos aspectos (ingreso per cápita familiar; fuentes de ingreso, propiedad y riqueza; empleo y educación; características de la vivienda; tener empleada doméstica). Luego construyen un índice unidimensional de bienestar: aplican PCA y toman el módulo del primer componente principal para proyectar los datos en una dirección de crecimiento del bienestar. A partir de este índice definen cuantiles multivariados, estableciendo un umbral inferior y superior que contiene a la clase media. Finalmente, también usan la estrategia de Fraiman, Justel & Svarc (2008) para seleccionar variables que permitan (i) reproducir con precisión el índice unidimensional, (ii) distinguir entre el grupo de pobres y el grupo de clase media vulnerable, (iii)

distinguir entre clase media vunerable y no vulnerable, (iv) distinguir entre clase media no vulnerable y clase alta. Encuentran que, en promedio en los distintos períodos de tiempo, las variables más relevantes para reproducir el índice unidimensional del bienestar son cuatro de las 19 originales: el pago en cuotas, el ingreso, el tipo de ocupación y tener personal doméstico. Concluyen que el bienestar es multidimensional: efectivamente el ingreso aparece como una variable relevante, pero otras dimensiones son importantes también. Además, el ingreso es menos confiable a la hora de distinguir grupos específicos.

Este trabajo explora el uso de las técnicas de PCA con pesos esparsos propuestas por Zou, Hastie & Tibshirani (2006) y Witten, Tibshirani & Hastie (2009) para seleccionar de forma no supervisada las variables relevantes que capturan la variación en el bienestar (la descripción de los métodos se encuentra en la sección 3). Los componentes obtenidos con PCA en general son combinaciones lineales de todas las variables originales, lo que en general dificulta la interpretación. Para el análisis de factores, Gasparini, Sosa-Escudero, Marchionni & Olivieri (2013) y Ferro Luzzi, Flückiger, & Weber (2008) facilitaban la interpretación de aplicando rotaciones que buscan que más variables tengan peso cero en los componentes. Sin embargo, las rotaciones no pueden garantizar que suficientes variables tengan peso cero en cada componente como para reconstruirlos sin tener que usar casi todas las variables originales. Es decir, no garantizan encontrar el mínimo subconjunto de K^* variables relevantes. En cambio, PCA con pesos esparsos combina linealmente sólo un subconjunto de las variables originales en cada componente, lo que sí permite determinar un subconjunto relevante de las variables originales y a la vez obtener componentes fácilmente interpretables. Otra alternativa sería establecer umbrales *ad hoc* que asignen un peso igual a cero a aquellas variables con valores absolutos de sus pesos menores a cierto valor en los componentes de PCA. Sn embargo, Cadima & Jolliffe (1995) argumentan que la interpretación con el enfoque de umbrales puede ser poco confiable y Zou, Hastie & Tibshirani (2006) muestran en simulaciones que el método de umbrales funciona peor que los métodos de PCA con pesos esparsos para identificar correctamente el subconjunto de variables que genera las variabilidad en los datos, y que alcanza una menor varianza explicada .

El objetivo principal de este trabajo, como en Caruso, Sosa-Escudero & Svarc (2015) y Edo, Sosa-Escudero & Svarc (2021), es identificar un subconjunto estricto de las variables originales. Ahora bien, a diferencia de las metodologías usadas en dichos trabajos, PCA con pesos esparsos permite seleccionar variables en el espacio original en un único paso. La selección se lleva a cabo de forma completamente no supervisada, sin construir una "etiqueta" previa a predecir, como lo es la pertenencia a determinado grupo de bienestar (i.e. pobres o clase media) en dichos trabajos previos. La relevancia de las variables estará asociada a su capacidad de explicar la variabilidad en los atributos de bienestar observados en los individuos, independientemente de una clasificación en grupos con mayores o menores niveles de bienestar. La metodología de PCA con pesos esparsos conserva una de las ventajas de PCA: la facilidad de interpretación, dado que las nuevas dimensiones son combinaciones lineales de las variables originales y a que provee de una cuantificación de la pérdida de la variabilidad explicada al dejar de considerar algunas de ellas. Finalmente, a diferencia de los métodos de clústers, PCA con pesos esparsos puede implementarse sobre un nuevo conjunto de datos usando la matriz de pesos estimada en el conjunto de datos inicial, sin la necesidad de entrenar de nuevo el algoritmo para cada conjunto de datos específico. Esto es porque PCA no sólo mapea las distancias existentes entre las observaciones en un conjunto dado de datos, sino que lo hace a partir de definir una estructura de relaciones entre las variables originales que es aplicable a cualquier otra muestra de las mismas variables.

La segunda contribución es usar escalamiento óptimo para cuantificar (i.e., asignar el número correspondiente a cada categoría) las variables categóricas. Las técnicas de escalamiento óptimo asignan un número

a cada una de las categorías optimizando algún criterio asociado al objetivo del análisis e imponiendo restricciones según el tipo de variable a transformar. En el caso de PCA, maximizando la suma de los primeros autovalores de la matriz de correlación de los datos cuantificados (Mair & De Leeuw, 2010). Para las variables ordinales, se impone la restricción de que la cuantificación respete el orden original entre las categorías, aunque las distancias entre ellas pueden variar (Mori, Koruda & Makino, 2016). Esto permite aprovechar casi todas las variables contenidas en la EPH, que son en su mayoría categóricas, e incluso considerar variables no necesariamente ordenables en términos de bienestar pero que dan cuenta de diferencias importantes entre grupos de población. A diferencia de los trabajos revisados, que parten de menos de 40 variables, aquí selecciono las relevantes en un conjunto de 126 variables iniciales, evitando una selección a priori respecto de qué variables conviene considerar, lo que suele tener cierto grado de arbitrariedad (Caruso, Sosa-Escudero & Svarc, 2015).

Merola & Baulch (2018) es el trabajo metodológicamente más cercano a este, dado que combina PCA con pesos esparsos y PCA con escalamiento óptimo. Además, hasta donde llega mi conocimiento, el único otro que aplica PCA con pesos esparsos en estudios de pobreza, desigualdad y desarrollo. En tanto, Coromaldi & Zoli (2012) usan escalamiento óptimo para construir indicadores multidimensionales de pobreza. A diferencia de este trabajo, el objetivo principal de Merola & Baulch (2018) no es medir la dimensionalidad del bienestar, sino construir un índice de activos a partir de la propiedad de 34 activos registrada en encuestas de hogares del noroeste de Vietnam y el noreste de Laos. Aunque son varios los trabajos en la intersección entre aprendizaje automático y estudios de pobreza, desigualdad y desarrollo que usan PCA para construir índices de activos o de riqueza que luego usan dentro de modelos predictivos (entre ellos Blumenstock, Cadamuro, & On, 2015, y Jean et al., 2016), muchos transforman las cuentas de activos en variables binarias antes de aplicar PCA, con lo que se pierde el ordenamiento original en la cantidad de activos. En cambio, Merola & Baulch (2018) aplican escalamiento óptimo para mantener el orden de las variables originales. Para validar los índices de activos construidos con esta metodología, analizan cuán buenos son para predecir el ranking de ingresos a nivel de hogar y a nivel per cápita, y encuentran que los índices de activos usando escalamiento óptimo mejoran la predicción del ranking de ingresos tanto a nivel hogar como a nivel per cápita. Por otra parte, los autores aplican técnicas de PCA con pesos esparsos para encontrar cuáles son los principales activos que explican la variación de la riqueza entre los hogares, e interpretar más fácilmente sus índices de activos. Merola & Baulch (2018) encuentran que al aplicar PCA con pesos esparsos los activos productivos pierden importancia en el primer componente principal, mientras que los bienes durables, la propiedad de medios de transporte y las características de la vivienda siguen siendo relevantes. Los índices construidos con componentes esparsos, usando entre un tercio y la mitad de los activos, llegan a resultados parecidos a PCA en las predicciones del ingreso.

3. Metodología

Parto de una matriz de datos X de dimensiones $N \times K$, donde K será el número de variables de la EPH y N el número de observaciones. Además, centro las variables de modo que las medias de las columnas de X sean iguales a cero y las escalo para que tengan varianza unitaria. El objetivo es encontrar un subconjunto $K^* \subset K$ relevante para explicar la variabilidad en X. Esto es, reducir la dimensión del espacio de variables de forma no supervisada (es decir, sin una variable dependiente asociada).

3.1. PCA

PCA es una método de aprendizaje no supervisado que permite reducir la dimensión de un conjunto de datos. Proyecta linealmente las K variables originales en un nuevo espacio de K componentes, que son las direcciones de máxima variabilidad en los datos en orden decreciente y ortogonales entre sí. Luego, es posible seleccionar un subconjunto de $P < K$ componentes principales que conformarán un espacio de menor dimensión que captura la mayor variabilidad posible de los datos originales. Los P componentes principales pueden ser pensados como los factores latentes que permiten generar la variación en los datos y cada uno de ellos resume un conjunto de variables originales correlacionadas (James et al., 2013).

Sea F la matriz de componentes, de dimensión $N \times K$ y cuya primer columna de F_1 es el primer componente principal. Las fila $i \in 1...N$ de F_1 indica el valor en el primer componente principal de la observación i. F es una combinación lineal óptima (en términos de varianza explicada) entre X y una matriz de pesos V de dimensión $K \times K$:

$$F = XV$$

La matriz V da cuenta del peso de cada variable original en cada componente principal, o en otras palabras, de la ponderación que corresponde a cada variable en la combinación lineal que conforma a cada componente. Se construye V de manera que F_1, el primer componente principal, tenga la mayor variabilidad posible. Si Σ es la matriz de covarianza de X y V_1^* la primer columna de V, entonces $Var(F_1) = {V_1^*}^T \Sigma V_1^*$. Por lo tanto:

$$V_1^* = \arg\max_{V_1} \quad V_1^T \Sigma V_1 \quad \text{s.a.} \quad V_1^T V_1 = 1$$

Donde V_1 son todos los posibles vectores de pesos. Recursivamente, podemos definir los siguientes componentes principales F_k con $k = \{2, .., K\}$ agregando la restricción de ortogonalidad:

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

Alternativamente, podemos obtener los componentes a partir de la descomposición de valor singular de X, donde la matriz diagonal D tiene como elementos a $d_1, ...d_K$ que son los autovalores de Σ ordenados de mayor a menor y V la matriz con los $V_1^*...V_K^*$ autovectores correspondientes:

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

La varianza del j-ésimo componente estará dada por el j-ésimo autovalor ordenado de Σ, $Var(F_j) = {V_j^*}^T \Sigma V_j^* = d_j$. Como los componentes no son correlacionados, la varianza total explicada por P componentes principales será la suma de los autovalores $\sum_{j=1}^{P} d_j = Tr(D)$, y la información que aporta cada componente (la proporción de variabilidad que explica) j está dada por $\frac{d_j}{Tr(D)}$.

3.2. Pesos esparsos

PCA tiene una desventaja: cada componente principal es una combinación lineal de las K variables y los pesos V son típicamente distintos de cero (Zou, Hastie & Tibshirani, 2006). Esto tiene dos consecuencias asociadas: dificulta la interpretabilidad de los componentes y significa que PCA en general no permite encontrar un subconjunto relevante K^* de las K variables originales. PCA permite reducir la dimensión en el sentido de obtener un nuevo espacio proyectado de menor dimensión, pero no reducir la dimensión en el espacio original de variables en el sentido de seleccionar un subconjunto de ellas.

En respuesta a este problema, los métodos que llamaré de "PCA con pesos esparsos" buscan reducir el número de variables con pesos distintos de cero en cada componente. Si algunas de las variables tienen peso igual a cero en el primer componente principal, existirá un $P < K$ tal que elegir P componentes principales permite explicar la mayor variabilidad posible con un subconjunto relevante K^* de las K variables originales. Esto es, se podrá seleccionar variables de forma no supervisada.

3.2.1. Regularización

Los métodos de pesos esparsos que consideraré hacen uso de técnicas de selección de variables (o regularización): LASSO (*Least Absolute Shrinkage and Selection Operator*) y su generalización en Elastic Net.

Sea un modelo de regresión lineal con N observaciones y K predictores, donde Y es el vector de respuesta y X la matriz con los predictores como columnas, que asumimos estandarizados. El estimador LASSO para el j-ésimo predictor puede obtenerse como:

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

Donde $\lambda \geq 0$ es una penalidad que se añade a mínimos cuadrados ordinarios que, si es lo suficientemente grande, hace que algunos de los coeficientes sean exactamente iguales a cero. El estimador Elastic Net, por su parte, añade una penalidad cuadrática dada por λ_2 (con $\lambda_2 \geq 0$) y puede obtenerse como:

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

En ambos casos, la esparsitud es consecuencia de la penalidad de norma 1 (λ), que genera soluciones de esquina a la hora de encontrar el vector de estimadores. Por su parte, la penalidad cuadrática (λ_2) es la propia de los estimadores Ridge, que pueden obtenerse como:

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

3.2.2. SPCA y SPC

Usaré dos métodos para encontrar componentes principales con pesos esparsos: *Sparse Principal Component Analysis* (SPCA), propuesto por Zou, Hastie & Tibshirani (2006) y *Sparse Principal Components* (SPC), propuesto por Witten, Tibshirani & Hastie (2009). Ambos métodos cuentan con algoritmos en librerías de R desarrolladas por los mismos autores, `elasticnet` y `PMA` respectivamente (Zou & Hastie, 2020; Witten &

Tibshirani, 2020). Giménez (2015) hace una revisión de otros métodos usados previamente en la literatura para encontrar pesos esparsos.

Sea nuevamente $X_{N\times K}$ la matriz de datos, $F_{N\times K}$ la matriz de componentes y $V_{K\times K}$ la matriz de pesos.

SPCA escribe el vector de pesos de PCA como el estimador resultante de un problema de regresión al cual se le puede introducir la penalidad de LASSO. LASSO permite obtener estimadores esparsos, y por lo tanto, pesos esparsos. Zou, Hastie & Tibshirani (2006) demuestran los siguientes pasos centrales a partir de los cuales es posible obtener SPCA:

1. Sea F_p el p-ésimo componente principal y V_p su vector de pesos. Dado que $F = XV$, es posible obtener V_p a partir de F_p y X. Sea $\lambda_2 \geq 0$ y

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

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

2. Agregando la penalidad de LASSO en (2), $\lambda \geq 0$,

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

   se obtiene una aproximación de V_p, $\hat{V}_p = \frac{\hat{\beta}}{||\hat{\beta}||}$. Con un λ lo suficientemente grande, se obtienen $\hat{\beta}$ esparsos, y por lo tanto pesos V_p esparsos.

3. Sin embargo, (3) requiere de conocer los componentes principales F para obtener la matriz de pesos V. Es posible modificar el problema para obtener ambos en simultáneo, resultando en lo que los autores llaman el "criterio SPCA" para obtener los P primeros componentes principales:

   Sea x_i el i-ésimo vector fila de la matriz X. Sea $A_{K\times P} = [\alpha_1, ..., \alpha_P]$, $B_{K\times P} = [\beta_1, ..., \beta_P]$ y $\lambda_p, \lambda_2 \geq 0$,

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

   entonces $\hat{\beta}_p \propto \hat{V}_p$ para $p = 1, ..., P$. Cuanto mayor la penalidad λ_p para el p-ésimo componente principal, más esparso será su vector de pesos.

Los autores además proponen un algoritmo iterativo para resolver (4), que requiere indicar el parámetro de esparsitud λ_p. Los autores sugieren probar con distintos λ_p y elegir aquel que permita un buen compromiso entre varianza capturada y esparsitud de cada componente.

Una diferencia con PCA es que SPCA no exige que los pesos sean ortogonales, y los componentes pueden estar correlacionados. Si F_p y F_{p-1} están correlacionados, la varianza del componente F_p puede contener las contribuciones a la varianza explicada de $F_1, ..., F_{p-1}$. Por lo tanto, los autores proponen una fórmula para ajustar el cálculo de la varianza total capturada por los primeros p componentes, usando la proyección lineal de F_p en el subespacio ortogonal al generado por $F_1, ..., F_{p-1}$. Es decir, regresando linealmente F_p en los componentes anteriores y tomando el residuo de dicha regresión para el cálculo de la varianza.

SPC es un caso particular de un método de descomposición penalizada de matrices que proponen Witten, Tibshirani & Hastie (2009), junto con un algoritmo iterativo para implementarlo. Sea u un vector columna de U de la descomposición de valor singular de X en (1), v un vector columna de V y $||w||_p$ la L_p-norma del vector w, es decir, $(\sum_i w_i^p)^{\frac{1}{p}}$. SPC puede obtenerse como:

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

Donde v es el vector de pesos del componente y c es el parámetro de esparsitud: acota la suma de valores absolutos de los pesos de v. Cuanto menor sea c, más esparso será v. Los autores proponen una estrategia similar a la validación cruzada para elegir el valor de este parámetro. Brevemente, consiste en los siguientes pasos. Para una grilla de valores posibles para c, construir 10 nuevas matrices a partir de X, llámense $X_1, ...X_{10}$, quitando a cada una de ellas un conjunto aleatorio disjunto del 10 % de las observaciones. Para cada $\hat{X}_m$ (donde $m \in 1, ..,10$) estimar el primer componente principal vía SPC, computar $\hat{X}_m = U_m D_m V_m$ y calcular el error cuadrático medio de estimar X_m usando $\hat{X}_m$. Luego tomar el promedio de los errores cuadráticos medios para $X_1, ...X_{10}$ y elegir el parámetro c que genere el menor promedio de errores cuadráticos medios. Este es un criterio basado en la estabilidad de las estimaciones, pero los autores señalan que si SPC se aplica con un objetivo de comprender mejor la estructura de los datos, otros criterios pueden ser más deseables, como por ejemplo alcanzar ciertos niveles deseables de esparsitud.

Como en SPCA, este método no garantiza la ortogonalidad entre v_k y v_{k-1}. En consecuecia, los autores proponen otro método para obtener una aproximación a la ortogonalidad, resolviendo para $k > 1$:

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

3.3. Escalamiento óptimo

PCA es una técnica pensada para variables numéricas, mientras que la EPH es un conjunto de datos de tipo mixto: tiene variables numéricas, pero la mayoría son categóricas, tanto nominales como ordinales. En general, los trabajos que usan este tipo de datos utilizan los códigos ad-hoc definidos para las categorías como números, o transforman las variables categóricas en variables binarias. Sin embargo, esto introduce restricciones innecesarias:

- Fija la misma distancia entre categorías contiguas, o entre cada categoría y la categoría base.
- En el caso de las variables nominales, también impone un orden creciente entre las categorías en un sentido arbitrario.
- En el caso de las variables ordinales, cuando se utilizan variables binarias se pierde el ordenamiento original entre categorías distintas a la base (Merola & Baulch, 2018)

Las técnicas de escalamiento óptimo, en cambio, transforman las variables categóricas en numéricas imponiendo las mínimas restricciones necesarias, y determinan la cuantificación (i.e., el número que corresponderá a cada categoría) de modo que se optimice algún criterio asociado al objetivo del análisis. En el caso de PCA, de modo que la cuantificación maximice la suma de los primeros P autovalores de la matriz de correlación de los datos cuantificados (Mair & De Leeuw, 2010).

Sea y_j el vector cualitativo con W_j categorías y N observaciones a cuantificar. Primero se codifica y_j usando una matriz indicadora $G_{N\times W_j}$, cuyos vectores columna son variables binarias. Sea el vector óptimamente escalado $y_j^* = G_j q_j$. El objetivo es encontrar q_j, el escalamiento óptimo, sujeto a restricciones según el tipo de y_j:

- Nominal: la cuantificación es irrestricta, solamente se garantiza que observaciones correspondientes a la misma categoría obtengan la misma cuantificación.
- Ordinal: cuantificación restringida al orden de las categorías. Si las categorías w_1 y w_2 son tales que $y_{jw_1} < y_{jw_2}$, entonces las categorías cuantificadas deben tener orden $y^*_{jw_1} < y^*_{jw_2}$.
- Numérico: ya cuantificado, simplemente se estandariza y_j para que tenga media cero y varianza unitaria.

PCA con escalamiento óptimo se denomina también "nonlinear PCA" (Mori, Koruda & Makino, 2016). Implemento el escalamiento óptimo de las variables categóricas con la librería de R `aspect` de Mair & De Leeuw (2018).

3.4. Modelos estimados

Se estimaron 9 modelos:

- PCA: modelo estimado usando PCA.
  1. `PCA`
- SPCA: modelos estimados con el método propuesto por Zou, Hastie & Tibshirani (2006), implementado en la librería `elasticnet` en R.
  2. `SPCA varnum`: modelo donde, de forma ad-hoc, impongo que el primer componente principal tenga solamente 15 variables con pesos distintos de cero; el segundo componente principal, 14 variables con pesos distintos de cero y así sucesivamente hasta el quinceavo componente.
  3. `SPCA lambda`: modelo donde elijo el parámetro de esparsitud λ que logre un buen compromiso entre varianza explicada y esparsitud de los primeros componentes. Como se explica en la sección 5.1.2, elijo $\lambda = 0{,}1$.
- SPC: modelos estimados con los métodos propuestos por Witten, Tibshirani & Hastie (2009), implementados en la librería `PMA` en R.
  4. `SPC`: modelo estimado con la metodología que no impone ortogonalidad de la ecuación (5) y con un parámetro de esparsitud $c = 6{,}4$ elegido con el método de validación cruzada que proponen Witten, Tibshirani & Hastie (2009) referido en la sección 3.
  5. `SPC esparso`: modelo estimado con la metodología que no impone ortogonalidad de la ecuación (5) y con un parámetro de esparsitud $c = 3{,}4$ que prioriza una mayor esparsitud, como se explica en la sección 5.1.2.
  6. `SPC ortog.`: modelo estimado con la metodología para obtener componentes aproximadamente ortogonales de la ecuación (6) y con un parámetro de esparsitud $c = 6{,}4$ elegido con el método de validación cruzada de Witten, Tibshirani & Hastie (2009) referido en la sección 3.

7. `SPC ortog. esparso`: modelo estimado con la metodología para obtener componentes aproximadamente ortogonales de la ecuación (6) y con un parámetro de esparsitud $c = 3{,}4$ que prioriza una mayor esparsitud, como se explica en la sección 5.1.2.

8. `SPC pos. ortog.`: modelo estimado con la metodología para obtener componentes aproximadamente ortogonales de la ecuación (6), con un parámetro de esparsitud $c = 5{,}2$, y que agrega la restricción de que los pesos de las variables en los componentes sean positivos.

9. `SPC pos. ortog. esparso`: modelo estimado con la metodología para obtener componentes aproximadamente ortogonales de la ecuación (6), con un parámetro de esparsitud $c = 3{,}2$ que prioriza una mayor esparsitud, y que agrega la restricción de que los pesos sean positivos.

3.5. Validación del método

Si el bienestar de la sociedad es aproximadamente estable en el tiempo -lo que es más probable cuanto más cercanos sean los períodos a comparar entre sí-, una forma de analizar la bondad de este método es ver si sus resultados son estables en el tiempo.

Primero, como Edo, Sosa-Escudero & Svarc (2020), verifico que los pesos asignados a cada variable en los primeros componentes principales sean estables entre encuestas de distintos momentos del tiempo.

Segundo, sea $F_{N\times P}$ la matriz con los valores que corresponden a cada individuo en cada uno de los P componentes principales. Dicho de otra forma, F es la matriz de las ubicaciones de cada individuo en medidas de bienestar construidas con P componentes principales. Es posible obtener F a partir de la matriz de pesos V y la matriz de datos originales X vía el producto matricial $F = XV$. Esto implica que es posible obtener la matriz de ubicaciones en t, F_t, de dos formas: usando la matriz de pesos obtenida tras aplicar alguno de los métodos de PCA a los datos de t, V_t, o usando la matriz de pesos obtenida a partir de los datos de $t-1$, V_{t-1}. Luego, comparo:

- Ubicación real $F_t = X_t V_t$
- Ubicación predicha $\hat{F}_t = X_t V_{t-1}$

Por un lado, comparo gráficamente las distribuciones de los valores de F_t y $\hat{F}_t$ para cada componente. Por el otro, construyo métricas de error. Sea (F_p) el p-ésimo componente principal. Se puede construir el error cuadrático medio (ECM) de predicción como:

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

Ahora bien, ECM_t es una medida de error que compara las ubicaciones absolutas de cada individuo en las medidas de bienestar. Sin embargo, puede resultar de mayor interés la ubicación relativa de los individuos en la medida de bienestar. Por lo tanto, comparo medidas de la ubicación relativa de los individuos en las medidas de bienestar F_t y $\hat{F}_t$. Hay dos casos posibles a la hora de medir la ubicación relativa de cada individuo en la medida de bienestar: haber tomado solamente el primer componente principal ($F_{N\times 1}$ y $\hat{F}_{N\times 1}$), o haber tomado $P > 1$ componentes principales ($F_{N\times P}$ y $\hat{F}_{N\times P}$).

3.5.1. Ubicación en el primer componente principal del bienestar

En el caso unidimensional es posible ubicar a cada individuo en los cuantiles de F_t y en los cuantiles $\hat{F}_t$. Luego, calcular el error cuadrático medio de clasificación en cuantiles reales y cuantiles predichos (que llamaré $ECMQ$), el porcentaje de individuos clasificados incorrectamente (el complemento de la *accuracy*, que llamaré *inacc*) y la máxima distancia entre el cuantil correcto y el incorrecto (*maxdist*). Sea Q_{it} el cuantil que corresponde al individuo i en F_t y $\hat{Q}_{it}$ el cuantil que corresponde al individuo i en $\hat{F}_t$, las métricas se calculan como:

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

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

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

Cuanto más cuantiles se consideren, más exigente será el criterio. En un extremo, se exigiría que el ordenamiento entre individuos sea exactamente igual en F y $\hat{F}$. Consideraré cuartiles y deciles.

3.5.2. Ubicación en el espacio de $P > 1$ componentes principales del bienestar

Para calcular las ubicaciones relativas de los individuos en un espacio de $P > 1$ componentes principales del bienestar como lo son $F_{N \times P}$ y $\hat{F}_{N \times P}$, una opción es recurrir a alguna noción de cuantil multivariado. Calcularé cuantiles de profundidad para cada individuo. La definición de un cuantil de profundidad central resuena a la definición de clase media adoptada por Gigliarano & Mosler (2009). En particular, considero dos medidas de profundidad:

1. Profundidad de Tukey: la profundidad del punto i será el mínimo número de puntos de cualquier semiespacio que contenga a i.

2. Distancia de Mahalanobis respecto de la mediana espacial: sea i_M la mediana espacial, aquel punto que minimice la suma de las distancias absolutas al resto de las observaciones. La distancia de Mahalanobis del punto i con respecto a i_M está dada por $\sqrt{(i - i_M)^T \Sigma^{-1} (I - I_M)}$, donde Σ es la matriz de covarianzas de X.

Una vez ubicado cada individuo en un cuantil de profundidad, pueden construirse las mismas métricas de error que para el caso de un solo componente principal: $ECMQ_t$, $inacc_t$ y $maxdist_t$.

Dependiendo de la distribución espacial de los datos, es posible que una medida de profundidad en torno a un centro no sea muy significativa para identificar grupos de individuos cercanos entre sí. En tal caso, una opción es usar algún método de clusterización, ubicar a cada individuo en un clúster de F_t y en un clúster de $\hat{F}_t$ y analizar qué porcentaje de los individuos son asignados al mismo clúster en ambos casos.

4. Datos

Se usaron los microdatos de las EPH a nivel individual y a nivel de hogar correspondientes a la región del Gran Buenos Aires en el tercer y cuarto trimestre de 2019 del Instituto Nacional de Estadísticas y Censos (INDEC, disponibles en www.indec.gob.ar). Dichos microdatos, así como la Clasificación Nacional de Ocupaciones (CNO) y la Clasificación de Actividades Económicas para Encuestas Sociodemográficas (CAES) fueron obtenidos usando la librería `eph` en R (Kozlowski et al., 2020). Para la elección del modelo y el análisis de la dimensionalidad del bienestar se usaron los datos correspondientes al cuarto trimestre de 2019, y para la validación se usaron también los del tercer trimestre de 2019. Luego del preprocesamiento detallado a continuación en 4.1, se obtuvieron dos bases de 126 variables, con 16525 para el tercer trimestre y 15610 observaciones para el cuarto. Además, cada variable puede clasificarse en distintos tipos de acuerdo a sus características, y fue óptimamente escalada, como se explicará en 4.2 y 4.3.

4.1. Preprocesamiento

Se tomaron las siguientes decisiones de preprocesamiento:

- *Nivel*. Se combinaron las encuestas a nivel individual y a nivel de hogar para construir una base donde cada observación corresponde a un individuo, pero que incluye tanto características individuales como características del hogar de cada individuo.
- *Orden en términos de bienestar*. Se incluyeron tanto variables que son ordenables en términos de bienestar (por ejemplo, el ingreso), como otras que no (por ejemplo, el rubro de actividad). Siempre que era posible hacerlo, se construyeron variables ordenables y se ordenaron las categorías originales de modo que valores más altos representen mayor bienestar.
- *Nuevas variables*. Se usaron muchas de las variables originales de la EPH, pero también se agregaron otras construidas a partir de ellas, como las usadas en Edo, Sosa-Escudero & Svarc (2020) y algunas variables que son parte de los Indicadores de Condiciones de Vida de los Hogares construidos por el INDEC (Gómez et al., 2004; INDEC, 2019). Esto permitió contar con más variables ordenables en términos de bienestar.
- Se distinguieron los casos de no-respuesta de aquellos casos donde no aplica una parte del formulario para la persona encuestada en cuestión (por ejemplo, cuando una persona empleada no responde preguntas de personas desempleadas).
- Se eliminaron las variables sin variabilidad y las perfectamente correlacionadas con otras variables, y se unificaron variables en los casos donde una de las variables podía ser incorporada como una categoría de otra. También se eliminaron unas pocas observaciones inconsistentes (por ejemplo, 5 observaciones cuyo estado de actividad indicaba que la persona era "desempleada" pero que en otra pregunta respondían que "no buscaban trabajo").
- *Nulos*: se eliminaron las observaciones con valores nulos, que corresponden a no-respuestas. Esto significa que el análisis solamente será representativo del conjunto de individuos que responden a todas las preguntas (aproximadamente el 55 % de las observaciones originales, con la mayoría de no-respuestas en variables de ingreso). A futuro podrían explorarse métodos de imputación.

4.2. Tipos de variables

Para facilitar el análisis es posible clasificar a variables en cuatro grupos temáticos: *Ingreso*, *Empleo*, *Habitacional* y *Miembros*. Este último grupo incluye otras características sociodemográficas como salud y educación. Además, podemos clasificar las variables por nivel (individual, hogar, vivienda), y según sean ordenables o no en términos de bienestar. La clasificación para cada variable en cada una de estas tres categorías se puede ver en el Anexo A, que también contiene una breve descripción de cada variable. En muchos casos, la definición puede consultarse en INDEC (2020).

4.3. Escalamiento óptimo

Las variables categóricas se cuantificaron (i.e., se asignó un número a cada categoría) con escalamiento óptimo, de modo que la cuantificación maximice el primer autovalor de la matriz de correlación de los datos transformados e imponiendo restricciones según el tipo de variable. Algunos ejemplos sobre cómo queda codificada cada tipo de variable:

- *Variables nominales.* Como no se impone ninguna restricción para el escalamiento, puede cambiar el orden y las distancias entre los números usados para codificar cada categoría. Por ejemplo, en el caso de la variable ESTADO, cuyas categorías originales eran 1 para ocupados, 2 para desocupados y 3 para inactivos, luego del escalamiento las categorías resultan -1.112 para ocupados, 0.905 para desocupados y 0.899 para inactivos. Esto significa que, para capturar mayor variabilidad en el primer componente principal, conviene ubicar a ocupados y desocupados como los más alejados entre sí, y a los inactivos como una categoría intermedia, pero mucho más cercana a los desempleados.
- *Variables ordinales.* Se impone la restricción de que la cuantificación mantenga el ordenamiento original entre las categorías, aunque las distancias entre ellas pueden variar. Por ejemplo, para la variable que indica la calidad de los materiales de los hogares, donde las categorías originales eran 0 para "insuficiente", 1 para "parcialmente insuficiente" y 2 para "suficiente", la cuantificación resulta en -2 para "insuficiente", -1.2 para "parcialmente insuficiente" y 0.6 para "suficiente". Se mantiene el orden entre las categorías, pero la distancia entre "suficiente" y "parcialmente insuficiente" es algo menor que entre "insuficiente" y "parcialmente insuficiente".
- *Variables numéricas.* Simplemente se estandarizan.

5. Resultados

Primero, se presenta una comparación entre los modelos estimados (detallados en 3.4) y se elige un criterio para elegir dos entre los modelos con pesos esparsos para el análisis posterior: `SPC ortog.` y `SPC ortog. esparso`. Segundo, se presenta el análisis de la dimensionalidad del bienestar en base los modelos elegidos. Tercero, se valida la metodología analizando la estabilidad de los resultados en el tiempo.

5.1. Elección de modelos con pesos esparsos

Esta subsección muestra los resultados de los modelos estimados usando PCA y PCA con pesos esparsos en términos de la varianza explicada por los primeros componentes y la esparsitud de los primeros componen-

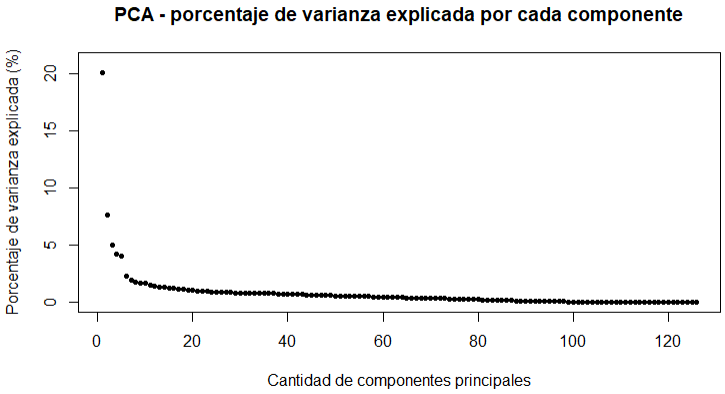

Figura 1: Varianza explicada por cada componente en `PCA`

tes, medida como cantidad de variables con peso igual a cero en cada uno de ellos. Se observa que en general hay un *trade-off* entre cuán esparsos son y cuánta varianza explican. Sin embargo, cada componente del modelo `SPC ortog.` explica prácticamente la misma variabilidad que `PCA` con ganancias importantes en esparsitud: usar la mitad de las variables para reconstruir el primer componente principal esparso, cuando reconstruir el primer componente principal de `PCA` requiere usarlas todas. En tanto, el modelo `SPC ortog. esparso` permite aún mayores ganancias en esparsitud si se toman sus 3 componentes principales esparsos para explicar la misma variabilidad que el primer componente principal de `PCA`: usar solamente el 35 % de las variables originales. Es decir, es posible capturar el 20 % de la varianza en el bienestar o bien con sólo el primer componente principal esparso de `SPC ortog.`, reconstruible con la mitad de las variables originales, o bien con los tres primeros componentes esparsos de `SPC ortog. esparso`, reconstruibles con sólo el 35 % de las variables originales.

5.1.1. PCA

La Figura 1 muestra la varianza explicada por cada uno de los componentes principales en `PCA`. El primer componente principal explica el 20 % de la varianza, una proporción importante teniendo en cuenta que se parte de 126 variables y notablemente más que el resto de los componentes (25 veces más que el promedio de la varianza explicada por cada componente). Como referencia, Edo, Sosa-Escudero y Svarc (2020) toman como índice de bienestar el primer componente principal que explica el 30 % de varianza partiendo de 19 variables, y Merola & Baulch (2018) toman como índice de activos el primer componente principal que explica el 21 % de la varianza partiendo de 34 activos en total. Además, se observa un quiebre a partir del sexto componente principal: los componentes del segundo al quinto explican entre el 7 % y el 4 % de la varianza, mientras que el sexto explica el 2 %. De todos modos, para el objetivo de resumir el bienestar en la menor cantidad de dimensiones posible, sería razonable conservar sólo el primer componente, que captura una varianza mucho mayor al resto.

La Figura 2 muestra que todas las variables tienen pesos distintos de cero en los dos primeros componentes de `PCA`. Esto dificulta su interpretación. Es cierto que algunas variables tienen mayor peso que otras en cada componente: en el primer componente principal se destaca ESTADO, que tiene un valor mucho más negativo que el resto, pero en valores absolutos no hay grandes diferencias. Las variables con mayores pesos son los deciles del ingreso de la ocupación principal. En tanto, en el segundo componente tienen mayor

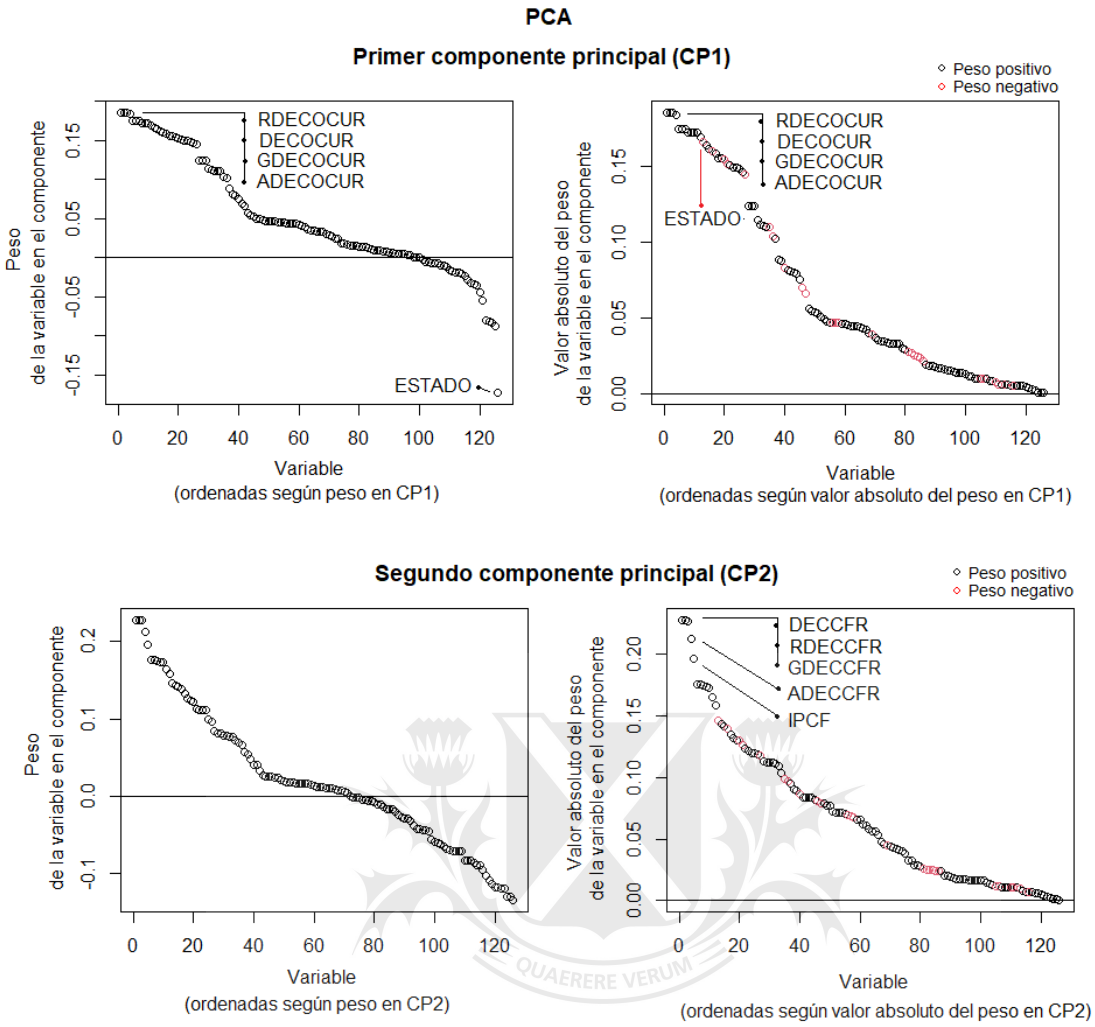

Figura 2: Pesos de cada variable en los dos primeros componentes principales en `PCA`

peso las variables de ingreso familiar. Ahora bien, no es posible seleccionar un subconjunto K^* de variables para construir los primeros componentes sin, por ejemplo, establecer umbrales *ad hoc* para asignar un peso igual a cero a aquellas variables con valores absolutos de sus pesos menores a cierto valor, lo que no es confiable en la práctica (ver sección 2). En cambio, recurrir a métodos de pesos esparsos permitirá obtener componentes interpretables -asociables a sólo un subconjunto de las variables originales- y determinar el mínimo conjunto de variables K^* que explique buena parte de la variabilidad en el bienestar.

5.1.2. Elección del parámetro de esparsitud

En los modelos SPCA y SPC es necesario definir el hiperparámetro de esparsitud: λ para SPCA y c para SPC (ver sección 3). La Figura 3 grafica la varianza capturada y la esparsitud de cada uno de los tres primeros componentes principales bajo distintos valores del hiperparámetro de esparsitud. El eje de ordenadas corresponde al porcentaje de varianza explicada por el componente para `SPCA lambda`, y

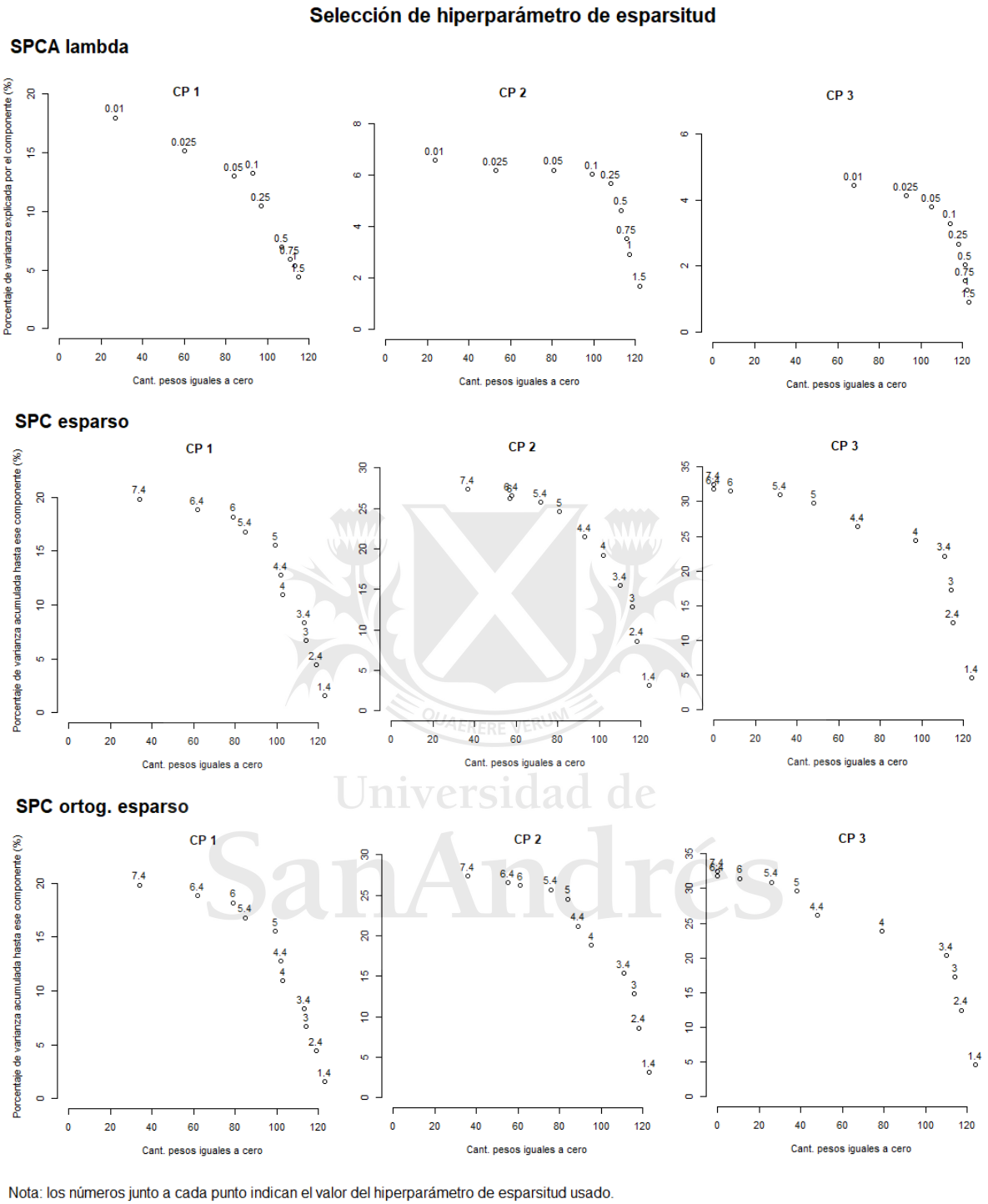

Figura 3: Elección del parámetro de esparsitud en modelos esparsos

al porcentaje de varianza explicada acumulado hasta el componente para `SPC esparso` y `SPC ortog. esparso`. El eje de abscisas mide la esparsitud como la cantidad de variables que tienen peso cero en el componente.

Se puede ver que hay un *trade-off* entre la variabilidad explicada por cada componente y su esparsitud: a medida que se eligen hiperparámetros de mayor esparsitud, la varianza explicada tiende a caer. La intensidad de dicho *trade-off* ya da una idea acerca de la multidimensionalidad del bienestar. Cuanto más pronunciado sea, significa que más variables del espacio original son necesarias para explicar buena parte de la variabilidad en el bienestar y por lo tanto, que el bienestar es "más" multidimensional. Si el bienestar fuera completamente unidimensional, podríamos encontrar un punto en el extremo superior-derecho: sería posible explicar la mayor parte de la varianza usando una sola variable original, y no existiría un conflicto entre varianza y esparsitud. Por el contrario, un *trade-off* más pronunciado se vería como una línea recta desde el extremo superior-izquierdo al extremo inferior-derecho: cualquier aumento de esparsitud traería una caída de la misma proporción en la varianza. En la Figura 3 se ve que existe un *trade-off* entre varianza y esparsitud, lo que sugiere que el bienestar es multidimensional, pero sin embargo hay puntos ubicados más hacia el extremo superior-derecho que otros, lo que sugiere que también hay espacio para reducir la dimensión. Los puntos que están ubicados más hacia el extremo superior-derecho indican hiperparámetros que permiten mayores ganancias de esparsitud con menores pérdidas de varianza explicada que otros hiperparámetros posibles.

Siguiendo a Zou, Hastie & Tibshirani (2006) elijo un valor para el hiperparámetro de esparsitud que logre un buen compromiso entre varianza explicada y esparsitud. Para `SPCA lambda` elijo $\lambda = 0{,}1$ porque permite aumentar la esparsitud del primer componente con incluso un mayor porcentaje de varianza explicada que un valor de $\lambda = 0{,}05$, y se ubica en el codo superior-derecho del segundo y tercer componente. En los modelos de SPC vemos que entre $c = 5$ y $c = 3{,}4$ hay una decisión conflictiva entre más varianza y más esparsitud. Dado que los modelos `SPC` y `SPC ortog.` priorizan la varianza y son poco esparsos (tienen un parámetro de esparsitud de $c = 6{,}4$, ver sección 3.4), elijo $c = 3{,}4$ para tener otro par de modelos que en cambio prioricen la esparsitud.

5.1.3. Comparación entre modelos

La Figura 4 compara los modelos estimados en términos de varianza explicada y esparsitud de cada uno de los primeros 15 componentes principales (CPs). El panel del centro muestra que `PCA` no es útil para determinar un subconjunto K^* de las variables originales que resuman la variabilidad en el bienestar: sea cual sea el porcentaje de varianza que queramos explicar (y la subsecuente cantidad de CPs que elijamos), siempre necesitaremos usar todas las variables, pues los pesos de todas ellas son distintos de cero en todos los CPs. Por lo tanto, es necesario recurrir a alguno de los modelos esparsos. ¿A cuál de todos ellos?

En general, también existe un *trade-off* entre los modelos en cuanto a la varianza explicada y la esparsitud de cada componente. Nuevamente, esto sugiere que el bienestar es multidimensional. Hay dos grandes grupos de modelos. Por un lado `SPC` y `SPC ortog.` priorizan la varianza: capturan prácticamente la misma varianza que `PCA` (panel superior), pero tienen una menor cantidad de variables con pesos iguales a cero que el resto de los modelos (panel central). Ahora bien, el hecho de que el primer componente de estos modelos capture la misma varianza que `PCA`, y a la vez la mitad de las variables tengan peso cero en él, da cuenta de que hay espacio para reducir la dimensión con componentes interpretables y que es posible capturar buena parte de la variabilidad en el bienestar con un subconjunto de las variables originales. Por el otro,

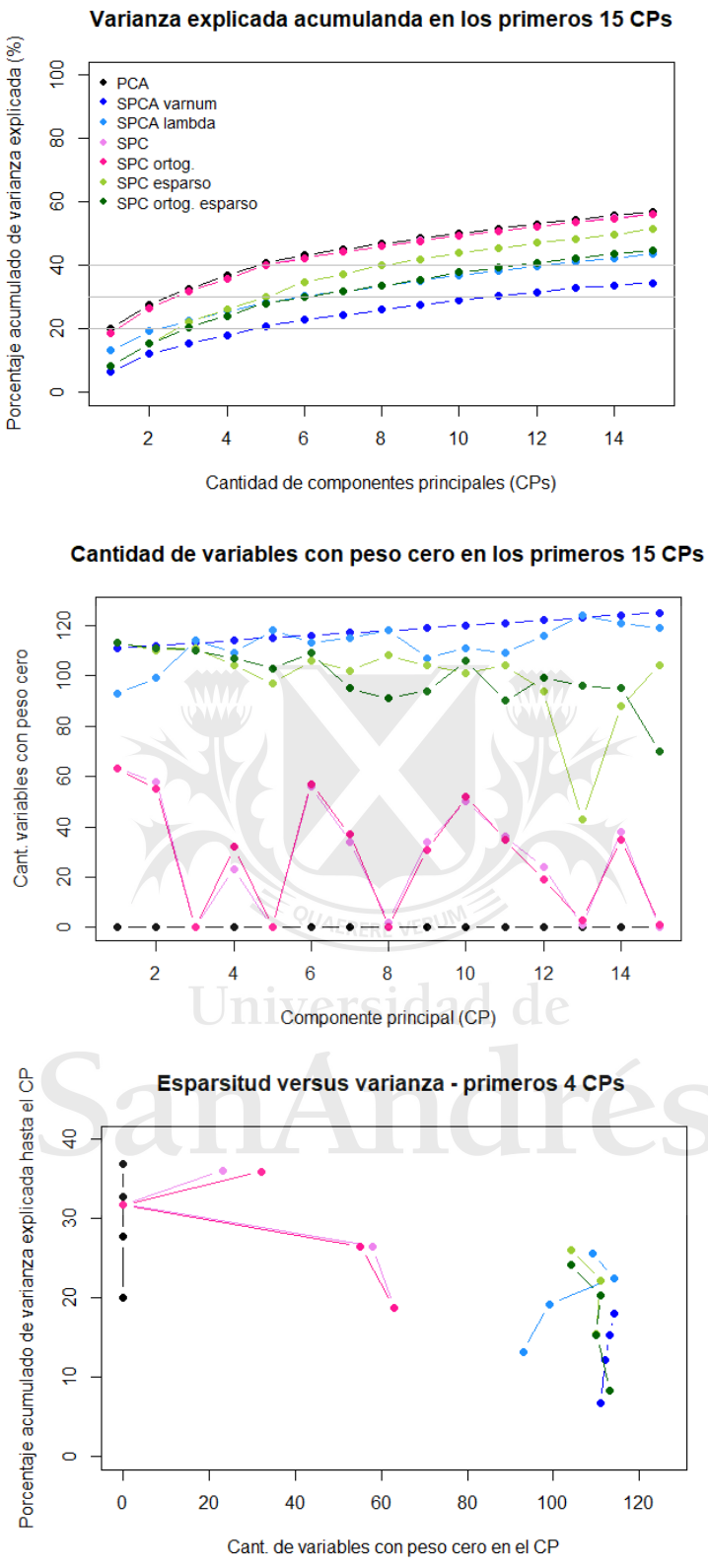

Figura 4: Comparación de modelos en términos de varianza explicada y esparsitud

SPCA y las versiones esparsas de SPC priorizan la esparsitud (panel central) con un costo en términos de varianza explicada (panel superior).

Entre los modelos estimados vía SPC, vemos que las versiones aproximadamente ortogonales son similares a las no ortogonales tanto en términos de varianza como de esparsitud en los primeros cinco componentes. De aquí en adelante, trabajaré con las versiones aproximadamente ortogonales ya que tienen ventajas en términos de interpretabilidad: aseguran que cada dimensión capture aspectos del bienestar poco correlacionados entre sí y por lo tanto permiten representar las observaciones y variables en biplots, analizar la correlación entre las variables y analizar las distancias entre los individuos en espacios de más de un componente principal sin cometer grandes distorsiones. Entre los modelos estimados vía SPCA, descarto `SPCA varnum` dado que explica muy poca varianza en relación al resto de los modelos. Entonces, tres modelos esparsos quedan preseleccionados: `SPC ortog.`, `SPCA lambda` y `SPC ortog. esparso`.

El Cuadro 1 muestra, para `PCA` y para cada uno de los modelos preseleccionados, y dado un nivel de varianza que deseemos explicar, cuántos componentes principales deberíamos considerar y cuántas variables $K^* \subset K$ necesitaríamos para reconstruir dichos componentes. El Cuadro 1 revela una forma alternativa de pensar el *trade-off* entre esparsitud y varianza: para un nivel dado de varianza explicada deseado, hay un *trade-off* entre cuán esparso es el modelo y la cantidad de componentes necesitamos para explicar dicha varianza. En otras palabras, hay un *trade-off* entre reducir las "dimensiones originales" (cuántas preguntas hacer en la encuesta) y reducir las "nuevas dimensiones" (cuántos componentes principales usar). Los tres modelos incluidos en el Cuadro 1 se ubican en distintos lugares del *trade-off*. Descarto `SPCA lambda` pues se encuentra en un lugar intermedio entre los otros dos, y no tiene la ventaja de ser aproximadamente ortogonal. De aquí en adelante, para el análisis usaré los modelos `SPC ortog.` y `SPC ortog. esparso`, que son complementarios entre sí:

- `SPC ortog.` captura la misma varianza que `PCA`, y por lo tanto sirve para resumir la variabilidad del bienestar en muy pocos componentes, pero a la vez ganar en esparsitud. **Permite construir una medida del bienestar usando solamente el primer componente principal** que explique el 20 % de la varianza (un porcentaje notable, dado que el punto de partida son 126 variables), y además construir ese componente **usando la mitad de las variables originales** (63 variables menos). Sin embargo, para explicar el 30 % de la variabilidad o más habría que usar todas las variables, dado que cada componente no es lo suficientemente esparso.
- `SPC ortog.esparso` permite explicar mayores porcentajes de varianza conservando la esparsitud (hasta un 30 % usando 50 variables menos, e incluso un 40 % descartando 11 variables). Es el modelo a usar si el objetivo principal es determinar el mínimo subconjunto K^* de las variables originales que expliquen buena parte de la variabilidad en el bienestar, por ejemplo para hacer encuestas más cortas. Sin embargo, requiere usar más cantidad de componentes: **3 componentes capturan el 20 % de la varianza usando 83 variables menos (el 35 % de las variables originales)**. Otra de sus ventajas es en términos de interpretabilidad: cada componente agrupa a un subconjunto muy pequeño de variables (en la Figura 4, panel central, se ve que los primeros tres componentes tienen más de 100 variables con peso igual a cero). Usaré este modelo para determinar un **subconjunto mínimo de variables** de la EPH para explicar el 20 % y 30 % de la variabilidad en el bienestar, y para interpretar la relevancia de distintos grupos de variables en el bienestar.

La Figura 5 muestra los pesos que `PCA`, `SPC ortog.` y `SPC ortog. esparso` asignan a cada variable en

Explicar varianza del...	PCA		SPC ortog.		SPCA lambda		SPC ortog. esparso	
	cant. CPs	cant. variables	cant. CPs	cant. variables	cant. CPs	cant. variables	cant. CPs	cant. variables
8 %							1	-113
13 %					1	-93	2	-98
20 %	1	todas	1	-63	2	-71	3	-83
25 %	2	todas	2	-37	4	-59	4	-64
30 %	3	todas	3	todas	6	-46	6	-50
40 %	5	todas	5	todas	12	-19	12	-11

Cuadro 1: Comparación de modelos en términos de varianza explicada y esparsitud

los dos primeros componentes principales. En los paneles superiores se ve que en los modelos esparsos se invierte el sentido de la proyección de las variables en el primer componente principal. Esto es, la misma variable que es ponderada en forma positiva en `PCA`, es ponderada de forma negativa en los modelos esparsos. Los paneles inferiores muestran los pesos en valor absoluto. Como la definición del sentido es arbitraria, en adelante multiplico por -1 los pesos de las variables en el primer componente de los modelos esparsos para facilitar la comparación. La Figura 5 muestra que, en general, los modelos más esparsos seleccionan aquellas variables con mayor peso en los modelos menos esparsos. Sin embargo, la selección no es uniforme. Es decir, hay variables seleccionadas (i.e. con peso distinto de cero en los modelos esparsos) que en `PCA` tienen un menor peso que otras variables que no son seleccionadas (i.e. tienen peso igual a cero en los modelos esparsos).

5.2. Las dimensiones del bienestar

La Figura 6 resume los modelos `PCA` y `SPC ortog.` en biplots. Los biplots grafican las proyecciones de las observaciones y las variables originales en los dos primeros componentes principales que, en estos dos modelos, capturan más del 25 % de la variabilidad en el bienestar. Los puntos representan los valores de las observaciones en los componentes principales, y los vectores representan los pesos de las variables en los componentes. Los puntos cercanos corresponden a individuos que tienen valores similares en los primeros dos componentes principales. Si dichos componentes representan adecuadamente los datos originales, también corresponden a individuos con valores similares en las variables originales, y puede decirse que son individuos similares en términos de bienestar. En tanto, los vectores que apuntan en la misma dirección corresponden a variables que tienen un perfil de respuesta en similar entre los individuos, y por ende, un significado parecido en términos de bienestar.

A partir de los biplots es posible hacer un primer análisis general respecto de qué tipo de variables captura cada dimensión. Las variables pueden ser clasificadas según (ver sección 4.2):

- Grupo temático (biplots en el panel superior). Se puede ver que en `PCA` y `SPC ortog.`, que son modelos poco esparsos, hay variables de todos los grupos en ambos componentes. En `SPC ortog.` se ve que las variables de empleo se asocian más bien al primer componente y las habitacionales al segundo.
- Nivel (biplots en el panel central). El primer componente se asocia más a variables a nivel individual; el segundo, a variables a nivel de hogar y vivienda.

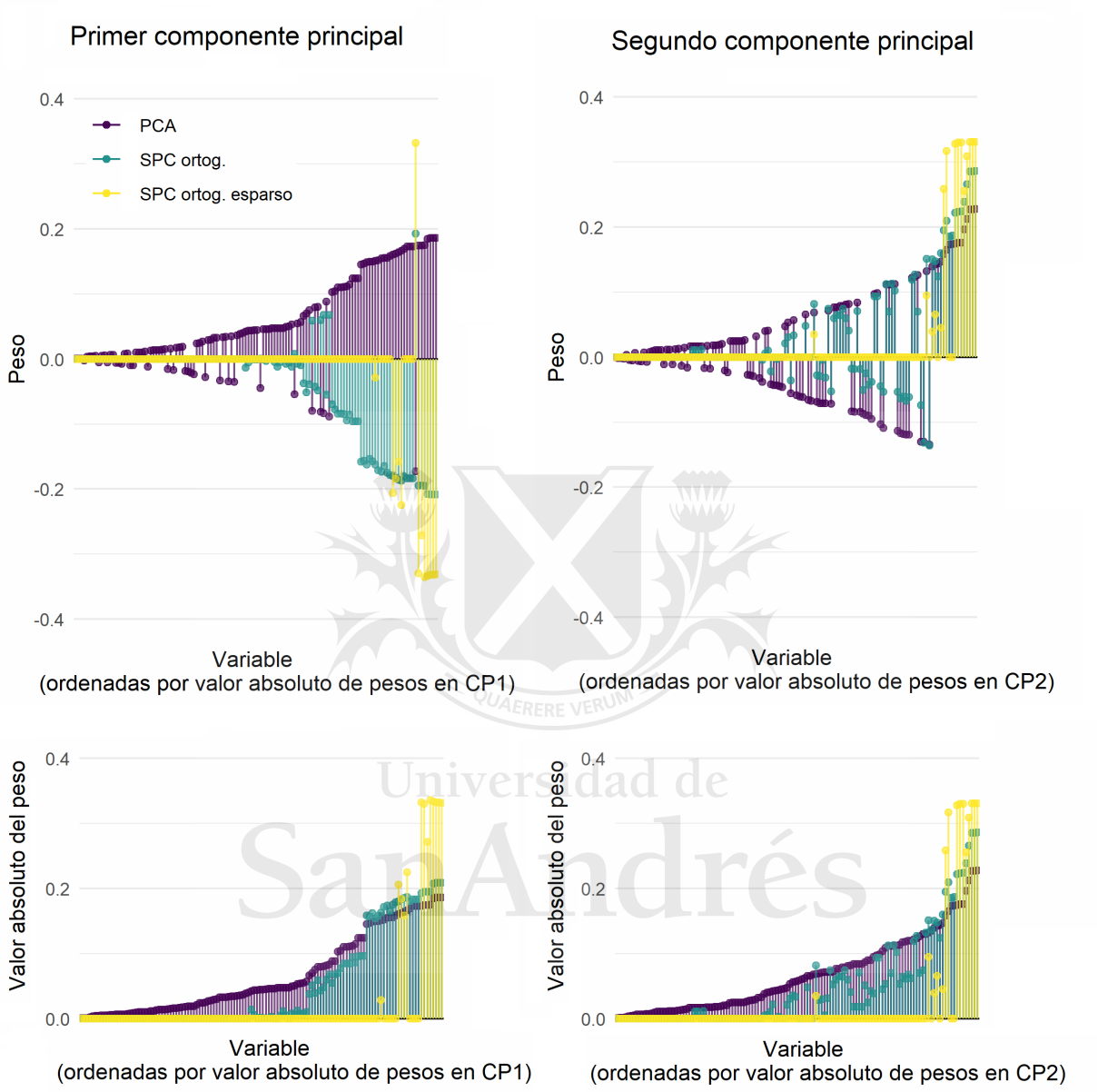

Figura 5: Variables seleccionadas en modelos esparsos

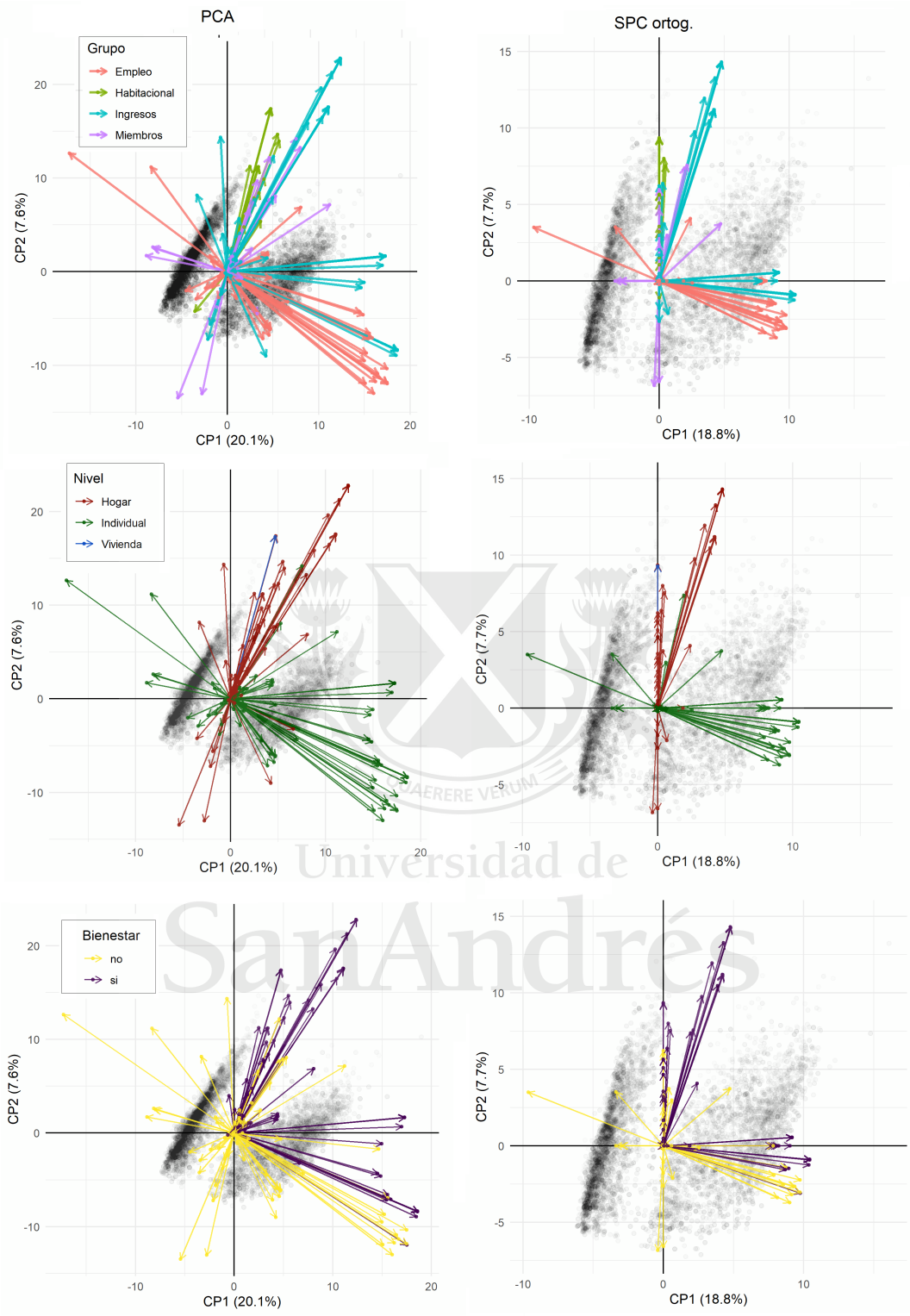

Figura 6: Biplots de `PCA` y `SPC ortog.`

- Si son ordenables o no en términos de bienestar (biplots en el panel inferior). En `SPC ortog.` es posible asociar mayores niveles de bienestar con valores más altos en cada componente.

Por otra parte, podemos ver qué variables tienen un perfil de respuesta en similar entre los individuos:

- Hay subgrupos de variables pertenecientes a los mismos temas que suelen estar correlacionadas entre sí. En el caso de los deciles de ingresos, por ejemplo, esto indica que quienes están en deciles superiores a nivel regional es posible que también estén en deciles superiores a nivel nacional. En el caso del empleo, podría estar reflejando la estructura jerárquica de la EPH. Además hay cercanía entre variables de distintos grupos temáticos, lo que significa que, por ejemplo, individuos con ciertas características de empleo suelen tener ciertos niveles de ingreso.
- El hecho de que las variables a nivel individual estén más asociados a un componente y las variables a nivel de hogar a otro, significa que los perfiles de respuesta a nivel hogar no se asocian a un único perfil de respuesta a nivel individual: dentro de un mismo hogar suele haber individuos con distintas características.
- La cercanía entre variables asociadas a un mayor bienestar significa que aquellos individuos con mayor bienestar en algunos aspectos, suelen tener mayor bienestar en otros, como se analizará en 5.2.1.

En los modelos más esparsos se mantiene una estructura de variables parecida a la de PCA. La esparsitud se refleja en que (i) menos variables están incluidas en los biplots de `SPC ortog` y (ii) los vectores que representan a las variables están más cercanos a los ejes, de modo que cada componente principal puede asociarse más inequívocamente a un subconjunto de variables. De hecho, si se quiere representar a `SPC ortog esparso` en un biplot, las variables quedan directamente superpuestas a los ejes, porque el subconjunto de variables con pesos distintos de cero en el primer componente tienen pesos iguales a cero en el segundo componente y viceversa, debido a la esparsitud.

Finalmente, se observa las observaciones se agrupan en dos grandes clusters, uno más disperso que otro. Los dos grupos visualizados con `PCA` tienen las siguientes características: un grupo más heterogéneo, mejor posicionado en promedio en términos de bienestar en variables individuales y peor posicionado en promedio en términos de bienestar en variables a nivel de hogar; y el grupo inverso (más homogéneo, mejor posicionado a nivel hogar y peor a nivel individual). Esta división es esperable si pensamos que una mejor posición a nivel hogar permite compensar, por ejemplo, un menor ingreso individual. Cuanto más esparso el modelo, la división entre clusters pasa a estar más determinada por el primer componente principal, es decir, asociados a diferencias a nivel individual, aunque dentro de cada uno de los grupos encontramos a individuos provenientes de hogares con distintas características.

5.2.1. Niveles de bienestar

La Figura 7 muestra los componentes que explican un 20 % de la varianza en los datos originales: el primer componente de `SPC ortog.` y los primeros tres componentes de `SPC ortog. esparso`. Grafica los pesos de las variables en cada uno de ellos, agrupándolas según si son ordenables o no en términos de bienestar. Las variables ordenables en términos de bienestar quedan todas orientadas hacia el mismo sentido (positivo) de los componentes, lo que significa que los individuos con mayor bienestar en ciertos aspectos, también suelen tener mayor bienestar en otros. Esto es esperable, dada las correlaciones registradas en la literatura entre distintos aspectos del bienestar (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013).

Figura 7: Variables asociables al bienestar en `SPC ortog.` y `SPC ortog. esparso`

¿Qué variables ordenables en términos de bienestar son las más relevantes para explicar su variabilidad? ¿Y a qué variables no ordenables *per se* a mayores niveles de bienestar están asociadas?

`SPC ortog. esparso` distingue tres direcciones principales de variación. En el primer componente, las variables asociadas positivamente al bienestar son los deciles del ingreso de la ocupación principal. Es interesante ver qué variables no ordenables (en todos los casos, variables categóricas nominales óptimamente escaladas) se asocian positivamente a esos mayores ingresos: estar empleado (ESTADO), trabajar en la Ciudad de Buenos Aires (PP09A); trabajar en oficinas/locales/establecimientos o vehículos (PP04G); el sector de actividad (caes) de los empleados o del último trabajo de los empleados, donde los ingresos individuales se asocian positivamente a trabajar en la administración pública o en servicios y negativamente a trabajar en servicio doméstico o en el sector de construcción; que tienen o tuvieron algún empleo, más probablemente en el sector público (PPA); ser patrón o asalariado (CAT); trabajar o haber trabajado en empresas grandes (C99); trabajar más horas en la ocupación principal (PP3E_TOT). En el segundo componente todas las variables son ordenables en términos de bienestar: las variables de ingreso familiar total y per cápita, en deciles y niveles, el nivel educativo del jefe de hogar, contar con cobertura médica (CH08), que el combustible usado para cocinar sea electricidad o gas (II8_ord), que los materiales de la vivienda sean suficientes (materiales) y que el jefe de hogar tenga una mayor calificación (jefe_CALIFICACION_ord). Finalmente, en el tercer componente las variables asociadas a mayores niveles de bienestar refieren al trabajo formal: mayores niveles de formalidad a la hora de cobrar (PP07K), que el trabajo sea por más tiempo (PP07D), tener beneficios como vacaciones pagas, aguinaldo, días por enfermedad y obra social (benef) y contar o haber contado con aportes jubilatorios (jub).

El primer componente de `SPC ortog.` podría usarse para aproximar un índice de bienestar dado que el sentido positivo se asocia a mayor bienestar, pero como incluye tanto variables ordenables como variables no ordenables en términos de bienestar, no es estrictamente monótono. Esto significa que distintas ubicaciones en dicho índice no necesariamente indican distintos niveles de bienestar. Por ejemplo, entre dos personas del mismo decil de ingresos, una persona inactiva tendrá un menor valor en el índice que una que está empleada y *per se* ser inactivo no implica tener menor bienestar que estar empleado. Lo que sucede es que ser inactivo está correlacionado negativamente a los deciles de ingreso de la ocupación principal, que sí son asociables a mayores niveles de bienestar. El énfasis de este trabajo no está puesto en construir un índice de niveles de bienestar (como sí lo está buena parte de la literatura relacionada, ver sección 2), aunque sí permite identificar grupos con distintas características en términos de bienestar, no necesariamente "mejores" o "peores". En ese sentido, un índice unidimensional construido con el primer componente de `SPC ortog.` o un índice multidimensional construido con los tres primeros componentes de `SPC ortog. esparso` permitiría distinguir un conjunto de individuos con características medianas de aquellos con características extremas, lo se asemeja a la definición de clase media de Gigliarano & Mosler (2009). Para construir un índice que permita ordenar a los individuos de mayores a menores niveles de bienestar había que tomar en consideración solamente las variables ordenables en términos de bienestar. Además, sería razonable cambiar la unidad de observación de individuos a hogares, o si no restringir la muestra eliminando a los menores. Luego, para poder asegurar que mayores valores del índice correspondan a mayor bienestar, hay dos opciones: aplicar como dirección de crecimiento el módulo del primer componente principal como en Edo, Sosa-Escudero & Svarc (2020) o restringir los pesos de SPC a ser mayores a cero, lo que puede tener como costo una menor variabilidad explicada en el primer componente (esto se observó al estimar los modelos `SPC pos. ortog.` y `SPC pos. ortog. esparso`).

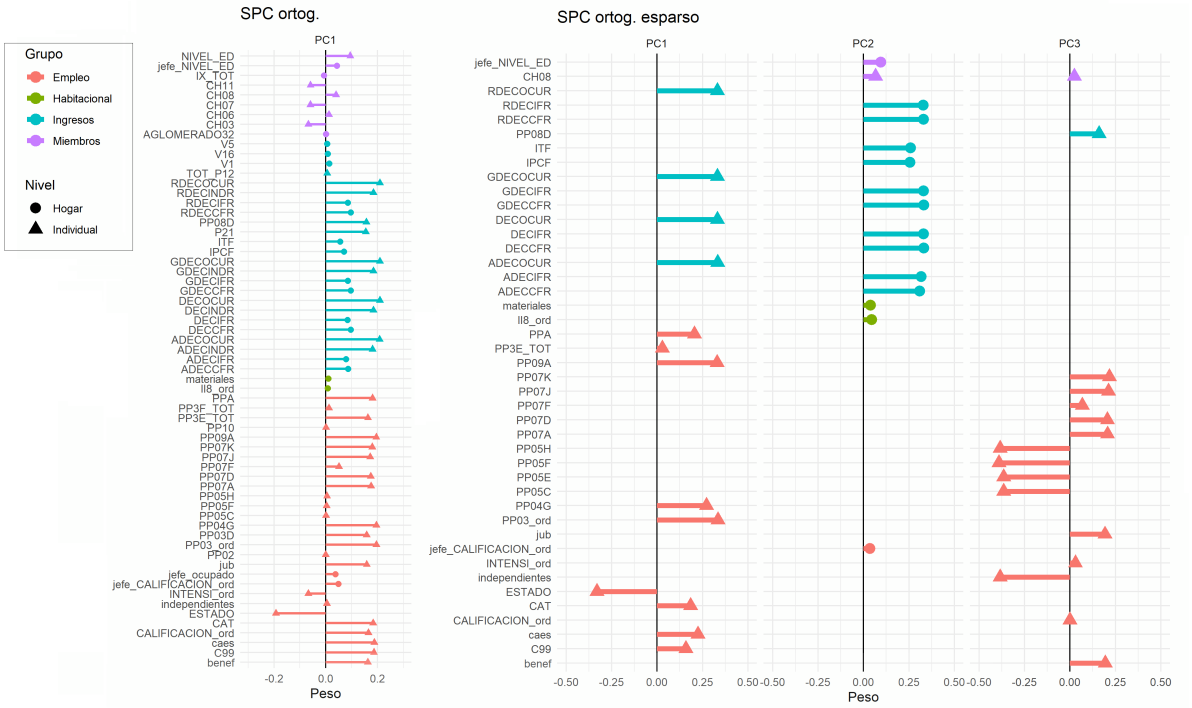

Figura 8: Variables relevantes por grupo y nivel en `SPC ortog.` y `SPC ortog. esparso`

5.2.2. Temas relevantes

Las variables más relevantes para explicar la variabilidad en el bienestar, ¿a qué aspectos del bienestar refieren? Para explicar el 20 % de la variabilidad, ya sea usando el primer componente principal de `SPC ortog.` o los primeros tres de `SPC ortog. esparso`, es necesario usar variables de los cuatro grupos temáticos (*Ingreso*, *Empleo*, *Habitacional*, y *Miembros*; ver sección 4.2). Esto indica que el bienestar es multidimensional, y no puede reducirse a un único grupo de variables. Ahora bien, el modelo `SPC ortog. esparso` permite ver qué grupos de variables resultan más relevantes y qué variables seleccionar dentro de cada grupo temático. El primer componente (8 % de la variabilidad) agrupa a los deciles de ingreso de la ocupación principal y variables de empleo (el estado de actividad, categoría ocupacional, sector y rubro de actividad, si trabaja en la Ciudad de Buenos Aires, tamaño de la empresa actual o anterior, en qué tipo de establecimiento realiza las tareas, horas trabajadas y no querer trabajar más horas). El segundo componente (7 % de la variabilidad) combina principalmente características del hogar: principalmente ingresos (deciles y niveles), pero también características sociodemográficas de los miembros del hogar (el nivel educativo del jefe y tener cobertura médica, la única variable a nivel individual con peso no nulo en el componente), características habitacionales (material de la vivienda y combustible para cocinar) y la calificación en el trabajo del jefe de hogar. Finalmente, el tercer componente (5 %) agrupa principalmente características de empleo, separando al grupo de los trabajadores independientes (si tienen una sociedad jurídica, cuántos clientes, si compra materias primas, entre otras) y de asalariados (preguntas asociadas a si tienen o tuvieron trabajo formal).

Es posible resumir cuáles son las variables más relevantes dentro de cada grupo temático:

- Ingresos: los deciles de la ocupación principal y los familiares. Para analizar cuán robusta es su im-

portancia, hice dos verificaciones. Primero, dejando solamente una de las variables de deciles (RDECOCUR), sigue siendo la más importante. Segundo, si los valores de RDECOCUR se permutan de forma aleatoria y sin reposición, deja de ser importante.

- Empleo: el estado de actividad, el tipo de actividad de los asalariados (tamaño de la empresa, sector y rubro, que esté ubicada en CABA, tipo de establecimiento, horas trabajadas), y la formalidad.
- Habitacional: los materiales de la vivienda y las fuentes de energía.
- Miembros: el nivel educativo del jefe de hogar y contar con cobertura de salud.

5.2.3. Encuestas más cortas

A continuación se resume el mínimo subconjunto de variables $K^* \subset K$ que se debe considerar para explicar un 20 % y un 30 % de la variabilidad en el bienestar usando `SPC ortog. esparso`. Los mínimos subconjuntos contienen solamente un 35 % y un 50 % de las variables originales respectivamente. Además, varias de las variables incluidas pueden construirse a partir de las mismas preguntas (por ejemplo, los deciles de ingreso), con lo cual el largo de la encuesta sería se reduciría aún más. Se listan las variables que es necesario ir agregando en cada componente adicional, sin volver a mencionar las ya listadas.

- **20 % de la varianza** en 3 componentes principales con 83 variables menos.
  1. Primer CP (8 % varianza acum.): PP03_ord (no querer trabajar más horas), ADECOCUR, GDECOCUR, DECOCUR (deciles de ingreso de la ocupación principal), ESTADO, RDECOCUR, PP09A (trabajar en la Ciudad de Buenos Aires), PP04G (establecimiento donde realiza las tareas), caes (rubro de actividad actual o anterior), PPA (sector), CAT (categoría de actividad o inactividad), C99 (tamaño de la empresa actual o anterior), PP3E_TOT (horas trabajadas en la ocupación principal).
  2. Segundo CP (15 % varianza acum.): DECCFR, GDECCFR, RDECCFR, DECIFR, GDECIFR, RDECIFR, ADECIFR, ADECCFR (deciles del ingreso del hogar), ITF, IPCF (ingresos del hogar), jefe_NIVEL_ED (nivel educativo del jefe), CH08 (cobertura de salud), II8_ord (combustible), materiales (condición habitacional), jefe_CALIFICACION_ord (calificación del jefe).
  3. Tercer CP (20 % varianza acum.): PP05F, PP05H, independientes, PP05C, PP05E (características empleo trabajadores independientes), PP07K, PP07J, PP07A, PP07D, benef, jub, PP07F (características del empleo asalariado), PP08D (ingresos asalariados), INTENSI_ord (sobre y subocupación), CH08 (contar con seguro de salud), CALIFICACION_ord (nivel de calificación).
- **30 % de la varianza** en 6 componentes principales con 64 variables menos. A las mencionadas anteriormente, deben agregarse:
  4. Cuarto CP (24 % varianza acum.): hacinamiento_v, hacinamiento, IX_TOT (habitantes por hogar y vivienda), CH06 (edad), W4 (porcentaje de miembros del hogar realizando tareas domésticas), usos (cuartos de de uso exclusivo respecto de los usados para dormir y trabajar), V1 (vivir de lo ganado en el trabajo), CH03 (relación de parentezco), V2_M (monto jubilación o pensión), IX_MEN10 (miembros menores a 10), CH11 (tipo de establecimiento educativo), W_11 (realizar las tareas domésticas), V2 (cobrar jubilación), jefe_ocupado (jefe de hogar ocupado), CH07 (estado civil).

	ECM_{4T}	Cuartiles	Deciles
Ubicación en el CP1 de `SPC ortog.`	0,00005	$ECMQ_{4T} = 0{,}001$ $inacc_{4T} = 0{,}1\,\%$ $maxdist_{4T} = 1$	$ECMQ_{4T} = 0{,}004$ $inacc_{4T} = 0{,}4\,\%$ $maxdist_{4T} = 1$

Cuadro 2: Métricas de error para la ubicación en el primer componente principal del bienestar

5. Quinto CP (28 % varianza acum.): PP10A, PP11LO, PP10, PP02, PP11L1 , PP11PQ, PP11ST, PP11R, PP11B1 (diferencias entre desocupados), PP07E11M (si trabaja o trabajó en período de empleo, prueba, pasantia), tiempo (que trabaja o trabajó), V3V4, V3V4_M (ingreso por indemnización o despido), NIVEL_ED (nivel educativo).
6. Sexto CP (30 % varianza acum.): ADECINDR, DECINDR, GDECINDR, RDECINDR (deciles del ingreso total individual), IX_TOT (cantidad de miembros en el hogar).

5.3. Validación

Si el bienestar es una característica estable en el tiempo de la sociedad, una forma de analizar la bondad de la metodología es ver si sus resultados son estables en el tiempo.

5.3.1. Pesos en el tiempo

Primero, comparo los pesos que los modelos `SPC ortog.` y `SPC ortog. esparso` asignan a cada variable según se estimen con la EPH correspondiente al cuarto trimestre de 2019 (4T2019) o al tercer trimestre de 2019 (3T2019). La Figura 9 muestra que, para `SPC ortog.` la coincidencia es casi exacta. Para `SPC ortog. esparso`, en general las variables con pesos distintos de cero son las mismas y sobre todo coinciden los pesos de las variables con pesos más altos en valor absoluto.

5.3.2. Ubicación en el nuevo subespacio de bienestar

Otra forma equivalente de cuantificar la estabilidad del método en el tiempo es comparando la ubicación de cada individuo en el nuevo subespacio de P componentes. Mido las diferencias entre:

- Ubicación real $F_{4T2019} = X_{4T2019}V_{4T2019}$: ubicación en el bienestar del cuarto trimestre calculada con los pesos del cuarto trimestre.
- Ubicación predicha $\hat{F}_{4T2019} = X_{4T2019}V_{3T2019}$: ubicación en el bienestar del cuarto trimestre calculada con los pesos del tercer trimestre.

La Figura 10 compara las distribuciones de las ubicaciones de los individuos en el primer componente de `PCA`, `SPC ortog.` y los primeros tres componentes de `SPC ortog. esparso`. `SPC ortog.` es más estable que `PCA`. El Cuadro 2 muestra las métricas de error para la predicción de la ubicación de los individuos en el primer componente principal de `SPC ortog.` descriptas en la sección 3.5. El error cuadrático medio de predicción es muy bajo, y mucho menor al de `PCA` (0,02). Por otra parte, la ubicación de los individuos en cuartiles o deciles del primer componente principal suele ser bastante precisa: sólo el 0,1 % y 0,4 % de los individuos son incorrectamente clasificados, y a lo sumo quedan clasificados en el cuantil aledaño.

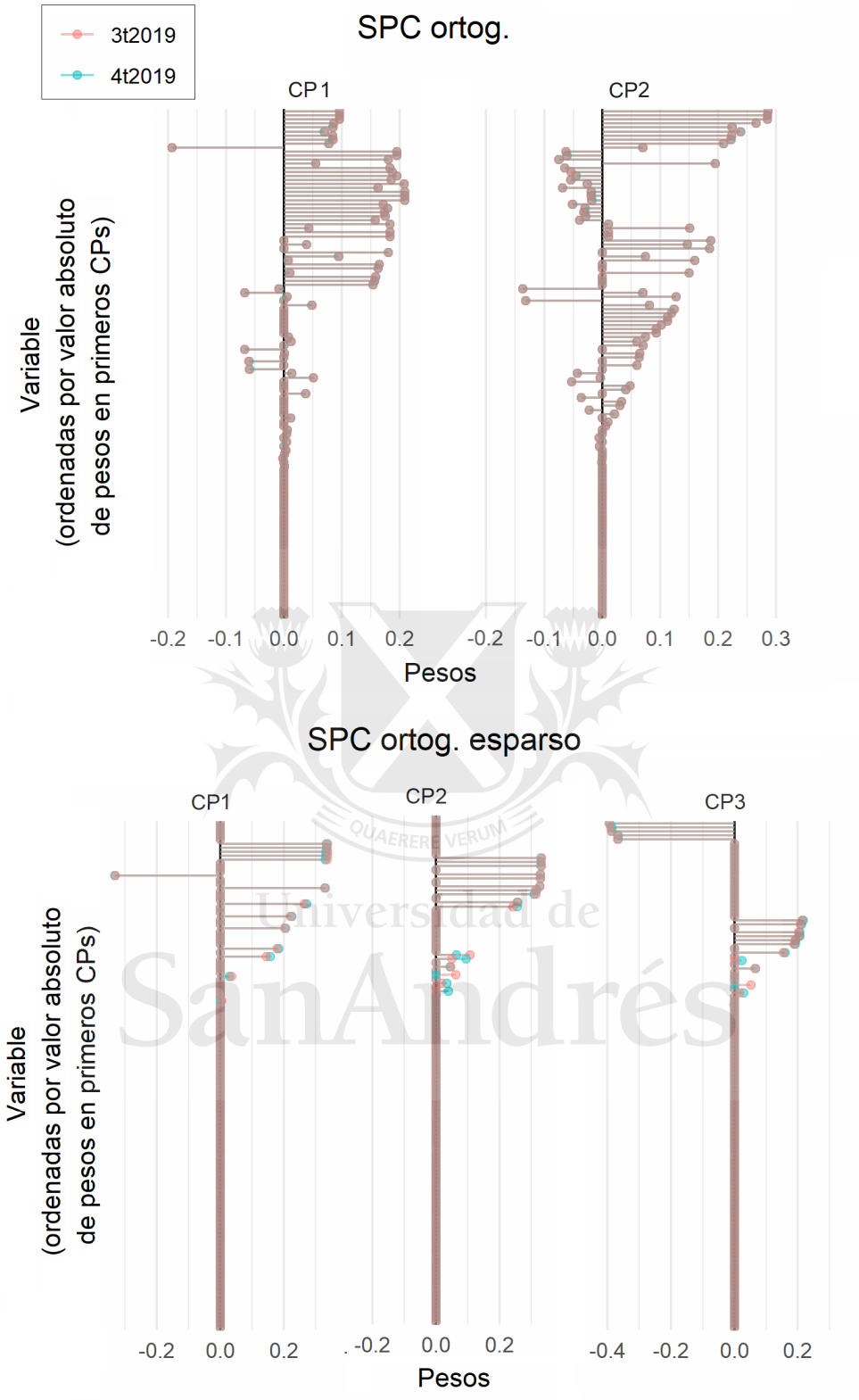

Figura 9: Estabilidad de los pesos en el tiempo

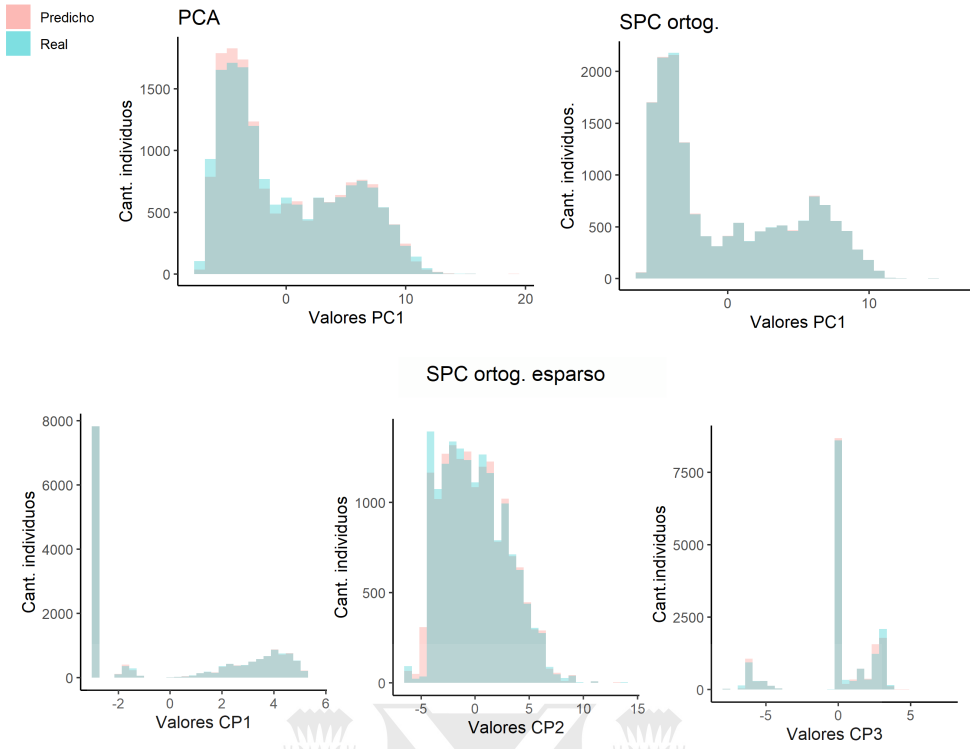

Figura 10: Estabilidad en el tiempo de las distribuciones de las ubicaciones de los individuos en los CPs

El Cuadro 3 muestra las métricas de error para la predicción de la ubicación de los individuos en el espacio de los tres primeros componentes principales de `SPC ortog. esparso`, también descriptas en la sección 3.5. Para calcular las medidas de profundidad eliminé observaciones duplicadas. Los paneles izquierdo y central de la Figura 11 grafican, para una submuestra aleatoria de 1000 individuos, la ubicación de cada uno de ellos en los cuartiles de profundidad. Se puede ver que una medida de centralidad y profundidades tomadas desde ella no tienen tanto sentido con la configuración espacial de los datos. De hecho, la clasificación en cuartiles con la profundidad de Tukey es bastante imprecisa, con un 14,9 % de las observaciones clasificadas incorrectamente y hasta con 3 cuartiles de diferencia. La distancia de Mahalanobis logra capturar uno de los clusters, de modo que gana algo de precisión y sólo clasifica incorrectamente el 3,5 % de los individuos, a lo sumo en cuartiles aledaños. Ahora bien, solamente el 0,7 % de las observaciones son clasificadas de forma incorrecta bajo ambas medidas de profundidad, lo que da cuenta de que la imprecisión en la clasificación en los cuartiles de profundidad con cada una de las medidas se debe a que no son medidas adecuadas para medir la ubicación relativa de los individuos en términos de bienestar con esta configuración espacial de datos. Por lo tanto, es preferible clasificar a los individuos usando algún método de clusters como k-medias, que permite encontrar razonablemente bien a los tres grupos, como se ve en el panel derecho de la Figura 11. La clasificación de los individuos en clusters hallados con k-medias es muy estable en el tiempo: sólo 17 individuos (el 0,1 %) quedan mal clasificados (Cuadro 3 y Cuadro 4).

En conclusión, los resultados son estables en el tiempo, tanto para la ubicación de los individuos en el primer componente principal esparso del bienestar estimado con `SPC ortog.`, como para la ubicación en el espacio de los tres componentes principales estimados con `SPC ortog. esparso`.

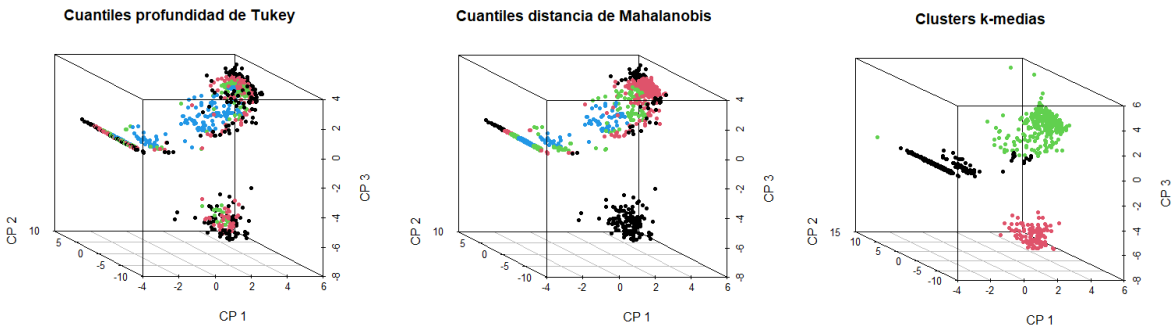

Figura 11: Medidas de ubicación en el espacio de 3 componentes principales de `SPC ortog. esparso`

	ECM_{4T}	Cuartiles profundidad de Tukey	Cuartiles distancia de Mahalanobis	Clusters k-medias
Ubicación en los primeros 3 CPs de SPC ortog. esparso	0,0042	$ECMQ_{4T} = 0{,}16$ $inacc_{4T} = 14{,}9\,\%$ $maxdist_{4T} = 3$	$ECMQ_{4T} = 0{,}035$ $inacc_{4T} = 3{,}5\,\%$ $maxdist_{4T} = 1$	$inacc_{4T} = 0{,}1\,\%$

Cuadro 3: Métricas de error para la ubicación en los primeros 3 CPs del bienestar

		Clúster predicho		
		1	**2**	**3**
Clúster real	**1**	8882	0	13
	2	0	1843	0
	3	4	0	4868

Cuadro 4: Clasificación en clusters obtenidos con k-medias

6. Conclusiones

Identificar el mínimo conjunto de variables necesarias para caracterizar al bienestar permitiría diseñar encuestas más cortas, más rápidas de implementar, menos costosas y con menos tasas de no respuesta. Este trabajo explora una metodología no aplicada hasta el momento para determinar la dimensionalidad del bienestar, según es representado en la EPH, e identificar el mínimo conjunto de variables necesarias para caracterizarlo: PCA con pesos esparsos. Compruebo que dicha metodología tiene dos ganancias respecto de PCA para este problema específico. Primero, permite explicar un porcentaje de variabilidad similar en el bienestar usando muchas menos variables originales. Segundo, es más interpretable: permite asociar cada dimensión relevante para el bienestar a un subconjunto de las variables originales. Por lo tanto, los métodos de PCA con pesos esparsos son útiles para explorar la hipótesis de multidimensionalidad del bienestar, esto es, cuántas dimensiones interpretables se requieren para caracterizarlo. Por otra parte, las técnicas de PCA no lineal permiten tomar como punto de partida 126 variables construidas con la EPH, que contiene variables de tipo mixto, en su mayoría categóricas.

El bienestar es multidimensional, aunque hay espacio para reducir la dimensión. La multidimensionalidad se refleja en el *trade-off* existente entre varianza explicada y esparsitud de los componentes que se observa en los distintos modelos estimados. Cuanto más pronunciado sea dicho *trade-off*, significa que más variables del espacio original son necesarias para explicar buena parte de la variabilidad en el bienestar y por lo tanto, que el bienestar es "más" multidimensional. Los resultados coinciden con la literatura previa: el ingreso es definitivamente importante, pero también son relevantes otras variables, fundamentalmente las asociadas al empleo. Para explicar el 20 % de la variabilidad en el bienestar se necesitan variables de distintos grupos temáticos: variables de ingreso, empleo, características habitacionales, educación y salud, tanto a nivel individual como del hogar. Ahora bien, es posible explicar ese 20 % de variabilidad usando sólo el 35 % de las variables originales, y el 30 % de la variabilidad usando solamente la mitad, lo que permitiría diseñar encuestas más cortas. De hecho, la cantidad de preguntas necesarias se reduce aún más, dado que varias de las variables seleccionadas pueden construirse a partir de la misma pregunta. Las variables más relevantes en cada grupo temático son los deciles del ingreso de la ocupación principal y del ingreso familiar; el estado de actividad y las características del tipo de actividad de los asalariados; los materiales y las fuentes de energía entre las características habitacionales; el nivel educativo del jefe hogar y tener cobertura de salud. La mayor variabilidad en el bienestar se asocia a tener o no altos ingresos en la ocupación principal, lo que correlaciona con estar empleado en cierto tipo de actividades (por ejemplo, importa la localización geográfica, el rubro de actividad y el tamaño de la empresa).

También en línea con la literatura previa, las variables ordenables en términos de bienestar están correlacionadas entre sí, es decir, los individuos con mayor bienestar en algunos aspectos suelen tener mayor bienestar en otros. Por otra parte, las variables a nivel individual aparecen asociadas al primer componente principal, mientras que las de hogar al segundo, lo que da cuenta de la heterogeneidad de características individuales dentro del hogar. Finalmente, el método es estable en el tiempo, tanto si se toma el primer componente principal esparso del bienestar, como si se toman los tres primeros.

Los resultados son relativos a encuestas similares a la EPH, que no necesariamente abarcan todas las dimensiones relevantes del bienestar. Por ejemplo, hay aspectos subjetivos que influyen en el bienestar y que no están reflejados en la encuesta (Gasparini, Sosa-Escudero, Marchionni & Olivieri, 2013). Determinar el conjunto de datos relevante del cual partir queda por fuera del alcance de este trabajo, y podría estudiarse a futuro. También se podría profundizar en otros aspectos. Primero, en elaborar índices que identifiquen

grupos de distintos niveles de bienestar. Este trabajo, en cambio, encuentra la variables relevantes para capturar la variabilidad en el bienestar independientemente de si son o no ordenables en términos de bienestar. Las características relevantes para capturar dicha variabilidad no tienen por qué coincidir con las que sirven para identificar a grupos de ricos o de pobres (Edo, Sosa-Escudero & Svarc, 2020). Con el objetivo de identificar niveles de bienestar, podrían llegar a ser útiles los métodos para seleccionar componentes principales con un objetivo supervisado (por ejemplo, *Lassoed PCA* de Witten & Tibshirani, 2008). Segundo, el hecho de que las características individuales y las características del hogar sean capturadas por componentes distintos podría ser útil para estudiar la movilidad social, observando movimientos en el tiempo en cada una de las direcciones. Finalmente, varias de las variables seleccionadas en los modelos esparsos están correlacionadas entre sí y se podría explorar cómo seleccionar sólo algunas de ellas. Una opción puede ser encontrar previamente los componentes principales esparsos de cada grupo de variables, lo que también podría servir para resumir mejor la estructura jerárquica de la EPH.

7. Referencias

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 y datos funcionales.* Tesis doctoral. Facultad de Ciencias Exactas y 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 y estructura para las bases preliminares Hogar y 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. Anexo: variables usadas

Nombre	Grupo	Nivel	Bienestar	Descripción
ADECCFR	Ingresos	Hogar	si	Ídem INDEC (2020)
ADECIFR	Ingresos	Hogar	si	Ídem INDEC (2020)
ADECINDR	Ingresos	Individual	si	Ídem INDEC (2020)
ADECOCUR	Ingresos	Individual	si	Ídem INDEC (2020)
AGLOMERADO32	Miembros	Hogar	no	Vive en Ciudad Autónoma de Buenos Aires
agua	Habitacional	Hogar	si	Acceso al agua (ver INDEC, 2019), usada en Edo, Sosa-Escudero & Svarc (2020)
ayuda	Ingresos	Hogar	no	Recibe ayuda en especie de personas o instituciones fuera del hogar
benef	Empleo	Individual	si	Si tiene días por vacaciones, aguinaldo, días por enfermedad, obra social (cuántos de esos)
C99	Empleo	Individual	no	Con cuantas personas trabaja actualmente o trabajó en su empleo anterior
caes	Empleo	Individual	no	Clasificador de Rama de Actividad
CALIFICACION_ord	Empleo	Individual	si	Calificación del empleo
CAT	Empleo	Individual	no	Agrupa CAT_OCUP y CAT_INAC de INDEC (2020)
CH03	Miembros	Individual	no	Ídem INDEC (2020)
CH04	Miembros	Individual	no	Ídem INDEC (2020)
CH06	Miembros	Individual	no	Ídem INDEC (2020)
CH07	Miembros	Individual	no	Ídem INDEC (2020)
CH08	Miembros	Individual	si	Ídem INDEC (2020)
CH09	Miembros	Individual	si	Ídem INDEC (2020)
CH10_ord	Miembros	Individual	si	Si asiste a establecimiento educativo si es menor a 17 años; si asiste o no asiste pero asistió si es mayor o igual a 17 años
CH11	Miembros	Individual	no	Ídem INDEC (2020)
CH15	Miembros	Individual	no	Ídem INDEC (2020)
CH16	Miembros	Individual	no	Ídem INDEC (2020)
DECCFR	Ingresos	Hogar	si	Ídem INDEC (2020)
DECIFR	Ingresos	Hogar	si	Ídem INDEC (2020)
DECINDR	Ingresos	Individual	si	Ídem INDEC (2020)
DECOCUR	Ingresos	Individual	si	Ídem INDEC (2020)
ESTADO	Empleo	Individual	no	Ídem INDEC (2020)
GDECCFR	Ingresos	Hogar	si	Ídem INDEC (2020)
GDECIFR	Ingresos	Hogar	si	Ídem INDEC (2020)
GDECINDR	Ingresos	Individual	si	Ídem INDEC (2020)
GDECOCUR	Ingresos	Individual	si	Ídem INDEC (2020)
hacinamiento	Habitacional	Hogar	si	Cantidad de habitaciones para uso exclusivo del hogar sobre el total de miembros del hogar
hacinamiento_v	Habitacional	Vivienda	si	Cantidad de habitaciones para uso exclusivo en la vivienda sobre total de miembros del hogar
II4	Habitacional	Hogar	si	Cantidad de habitaciones que son cocina, lavadero o garage
II7	Habitacional	Hogar	no	Ídem INDEC (2020)
II7_propietario	Ingresos	Hogar	si	Es propietario de vivienda y/o terreno, usada en Edo, Sosa-Escudero

Nombre	Grupo	Nivel	Bienestar	Descripción
				& Svarc (2020)
II8_ord	Habitacional	Hogar	si	Combustible usado para cocinar: electricidad o gas de red para cocinar; gas de tubo o garrafa; kerosene, leña o carbón
II9_ord	Habitacional	Hogar	si	Ídem INDEC (2020)
independientes	Empleo	Individual	no	Tipo de empresa: sin socios, sociedad jurídica, otra sociedad familiar, otra sociedad no familiar.
INTENSI_ord	Empleo	Individual	no	Si no es sobreocupado o subocupado por insuficiencia horaria
IPCF	Ingresos	Hogar	si	Ídem INDEC (2020)
ITF	Ingresos	Hogar	si	Ídem INDEC (2020)
IV1	Habitacional	Hogar	no	Tipo de vivienda
IV1_ord	Habitacional	Hogar	si	Vive en casa o departamento
IV12	Habitacional	Hogar	si	Ubicación en área riesgosa, usada en Edo, Sosa-Escudero & Svarc (2020)
IX_MEN10	Miembros	Hogar	no	Ídem INDEC (2020)
IX_TOT	Miembros	Hogar	no	Ídem INDEC (2020)
jefe_CALIFICACION_ord	Empleo	Hogar	si	Jefe de hogar calificado, usada en Edo, Sosa-Escudero & Svarc (2020)
jefe_CH09	Miembros	Hogar	si	Jefe de hogar alfabetizado, usada en Edo, Sosa-Escudero & Svarc (2020)
jefe_NIVEL_ED	Miembros	Hogar	si	Nivel educativo del jefe de hogar, usada en Edo, Sosa-Escudero & Svarc (2020)
jefe_ocupado	Empleo	Hogar	si	Jefe de hogar ocupado, usada en Edo, Sosa-Escudero & Svarc (2020)
jub	Empleo	Individual	si	Si en el trabajo actual (o en el último trabajo si es desempleado) tiene descuento jubilatorio, aporta por si mismo o no aporta.
materiales	Habitacional	Hogar	si	Índice de privación material de los hogares (ver INDEC, 2019), usada en Edo, Sosa-Escudero & Svarc (2020)
NIVEL_ED	Miembros	Individual	no	Ídem INDEC (2020)
P21	Ingresos	Individual	si	Ídem INDEC (2020)
PP02	Empleo	Individual	no	Agrupa PP02C y PP02E de INDEC (2020)
PP02H	Empleo	Individual	no	Ídem INDEC (2020)
PP02I	Empleo	Individual	no	Ídem INDEC (2020)
PP03_ord	Empleo	Individual	si	Agrupa PP03G, PP03GH, PP03GI y PP03GJ de INDEC (2020): si no quería trabajar más horas (en la semana, en el mes, buscar otro empleo ocupación o actividad), si quería y podía, si quería pero no podía, etc.
PP03D	Empleo	Individual	no	Ídem INDEC (2020)
PP04B2	Empleo	Individual	no	Ídem INDEC (2020)
PP04G	Empleo	Individual	no	Ídem INDEC (2020)
PP05C	Empleo	Individual	no	Si tiene maquinaria, local o vehiculo propio
PP05E	Empleo	Individual	no	Ídem INDEC (2020)
PP05F	Empleo	Individual	no	Ídem INDEC (2020)
PP05H	Empleo	Individual	no	Ídem INDEC (2020)
PP07A	Empleo	Individual	no	Ídem INDEC (2020)
PP07D	Empleo	Individual	si	Ídem INDEC (2020)

Nombre	Grupo	Nivel	Bienestar	Descripción
PP07E11M	Empleo	Individual	no	Agrupa PP07E y PP11M de INDEC (2020)
PP07F	Empleo	Individual	no	Si recibe comida gratis, vivienda, algún producto o servicio, otro beneficio (cuántos de esos)
PP07J	Empleo	Individual	no	Ídem INDEC (2020)
PP07K	Empleo	Individual	si	Ídem INDEC (2020)
PP08D	Ingresos	Individual	no	Agrupa PP08D1 y PP0D4 de INDEC (2020)
PP08F	Ingresos	Individual	no	Agrupa PP08F1 y PP08F2 de INDEC (2020)
PP08J1	Ingresos	Individual	si	Ídem INDEC (2020)
PP08J2	Ingresos	Individual	no	Ídem INDEC (2020)
PP08J3	Ingresos	Individual	no	Ídem INDEC (2020)
PP09A	Empleo	Individual	no	Ídem INDEC (2020)
PP10	Empleo	Individual	si	Ídem INDEC (2020)
PP10A	Empleo	Individual	si	Ídem INDEC (2020)
PP11B1	Empleo	Individual	no	Ídem INDEC (2020)
PP11L1	Empleo	Individual	no	Ídem INDEC (2020)
PP11LO	Empleo	Individual	no	Agrupa PP11L y PP11O de INDEC (2020)
PP11PQ	Empleo	Individual	no	Agrupa PP11P y PP11Q de INDEC (2020)
PP11R	Empleo	Individual	no	Ídem INDEC (2020)
PP11ST	Empleo	Individual	si	Agrupa PP11S y PP11T de INDEC (2020)
PP3E_TOT	Empleo	Individual	no	Ídem INDEC (2020)
PP3F_TOT	Empleo	Individual	no	Ídem INDEC (2020)
PPA	Empleo	Individual	no	Agrupa PP04A y PP11A de INDEC (2020)
RDECCFR	Ingresos	Hogar	si	Ídem INDEC (2020)
RDECIFR	Ingresos	Hogar	si	Ídem INDEC (2020)
RDECINDR	Ingresos	Individual	si	Ídem INDEC (2020)
RDECOCUR	Ingresos	Individual	si	Ídem INDEC (2020)
saneamiento	Habitacional	Hogar	si	Acceso al saneamiento (ver INDEC, 2019), usada en Edo, Sosa-Escudero & Svarc (2020)
tiempo	Empleo	Individual	no	Hace cuánto tiempo que trabaja allí
TOT_P12	Ingresos	Individual	si	Ídem INDEC (2020)
usos	Habitacional	Hogar	si	Cuartos totales de uso exclusivo respecto de los usados para dormir y para trabajar
usos_II4	Habitacional	Hogar	si	Cantidad de habitaciones que son cocina, lavadero o garage que no se usan para dormir
V1	Ingresos	Hogar	no	Vivió de lo que ganan en el trabajo
V10	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V10_M	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V11_M	Ingresos	Hogar	no	Ídem INDEC (2020)
V12_M	Ingresos	Hogar	no	Ídem INDEC (2020)
V13	Ingresos	Hogar	no	Ídem INDEC (2020)
V14	Ingresos	Hogar	no	Ídem INDEC (2020)
V15	Ingresos	Hogar	no	Ídem INDEC (2020)
V16	Ingresos	Hogar	no	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V17_ord	Ingresos	Hogar	si	Ídem INDEC (2020)
V18_M	Ingresos	Hogar	no	Ídem INDEC (2020)
V2	Ingresos	Hogar	no	Vivio de alguna jubilación o pensión (o retroactivo) cobrada este mes o el anterior
V2_M	Ingresos	Hogar	no	Ídem INDEC (2020)
V21_M	Ingresos	Hogar	si	Ídem INDEC (2020)
V3V4	Ingresos	Hogar	no	Vivió de indemnización por despido o seguro

Nombre	Grupo	Nivel	Bienestar	Descripción
				de desempleo
V3V4_M	Ingresos	Hogar	no	Monto recibido por indemnización por despido y seguro de desempleo
V5	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V5_M	Ingresos	Hogar	no	Ídem INDEC (2020)
V8	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V8_M	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V9	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
V9_M	Ingresos	Hogar	si	Ídem INDEC (2020), usada en Edo, Sosa-Escudero & Svarc (2020)
W_11	Miembros	Individual	no	Realiza las tareas del hogar
W_21	Miembros	Individual	no	Ayuda en las tareas del hogar
W_domestico	Miembros	Hogar	si	El hogar tiene empleada doméstica, usada en Edo, Sosa-Escudero & Svarc (2020)
W_externo	Miembros	Hogar	no	Alguien externo realiza o ayuda en tareas del hogar
W4	Miembros	Hogar	no	Cantidad de personas que realizan tareas del hogar sobre el total de miembros del hogar
W5	Miembros	Hogar	no	Cantidad de personas que ayudan en tareas del hogar sobre el total de miembros del hogar