Índices Socioeconómicos con Capacidad Predictiva vía Reducción de Dimensiones Supervisada

Resumen
La obtención de índices confiables que reflejen el nivel socioeconómico de la población resulta fundamental para el diagnostico de la realidad social y para la planificación de políticas públicas que tengan como objetivo la reducción de la pobreza. Un objetivo práctico de los índices de estatus socioeconómico (*SES*) es poder contar con un predictor de otras variables económicas y sociales, como ser la pobreza o ingreso monetario, cuando dicha información, por diversos motivos, no se encuentra del todo disponible. Entre las metodologías estadísticas para la construcción del índices ESE predominan las no supervisadas, como ser Componentes Principales y sus extensiones para variables no continuas. Sin embargo, tales métodos no aprovechan la información contenida en alguna respuesta de interés, perdiendo con ello, poder predictivo. En el presente trabajo se muestra cómo el método desarrollado recientemente por Forzani et al. (2018), basado en el enfoque de Reducción Suficiente de Dimesiones, se aplica a la construcción de índices *SES* para predecir ingreso y pobreza monetaria. Adicionalmente, se muestra cómo esta técnica supervisada supera, en varios sentidos, a las otras metodologías no supervisadas que se usan corrientemente.

Palabras Clave: Reducción Suficiente de Dimensiones; Variables Ordinales; Componentes Principales; Pobreza.

Predictive Socioeconomic Status indices with Supervised Dimension Reduction

Abstract
Obtaining reliable indices that reflect the socioeconomic level of the population is crucial for the diagnosis of social reality and for the planning of public policies that aim to reduce poverty. A practical objective of the socioeconomic status (*SES*)index is to have a predictor of other economic and social variables, such as poverty or monetary income, when this information, for various reasons, is not completely available. Among the statistical methodologies for the construction of the SES indices, predominate the non-supervised reduction methods, such as Principal Components Analysis (PCA) and their extensions for non-continuous variables. However, such methods do not take account the information contained in a response varibale of interest, thus losing predictive power. This paper shows how the method recently developed by Forzani et al. (2018), based on the Sufficient Dimension Reduction approach, is applied to the construction of *SES*indexes to predict income and monetary poverty. Additionally, it shows how this supervised technique perform better than the other unsupervised methodologies that are currently used in empirical research.

Keywords: Sufficient Dimension Reduction; Ordinal variables; Pricipal Components; Poverty.

1. Introducción

En ciencias sociales, diferentes medidas del nivel o estatus socio-económico de los agentes son ampliamente utilizadas como predictoras de distintos comportamientos individuales o de resultados sociales (ej. Roy & Chaudhuri (2009), Murasko (2009), Kamakura & Mazzon (2013), Mazzonna (2014), Feeny et al. (2014), García Arancibia et al. (2015)). Para los gobiernos y organizaciones no gubernamentales que llevan adelante diferentes políticas o programas de protección social, resulta crucial la predicción del nivel socio-económico de un hogar a través de la observación de variables simples que lo caracterizan (Mokomane (2013), Richardson & Bradshaw (2012)). Desde esta manera, los índices socio-económicos (comúnmente denominado índice SES: *Socio-Economic Status index*) tienen como objetivo predecir o describir fenómenos sociales tales como el ingreso, la pobreza o cuestiones relacionadas con la salud (sintetizados en una variable respuesta Y), utilizando para su construcción variables indirectas (i.e. predictoras $\mathbf{X}$), generalmente de naturaleza categórica, que son más fáciles de obtener y de medir que la variable de interés a predecir Y.

La noción de estatus socio-económico contempla diversas dimensiones, incluyendo el ingreso monetario, la riqueza (e.g. en activos), el nivel educativo alcanzado, el tipo de empleo u ocupación y otras aspectos que muestran una cierta estratificación acerca de la posición económica o social de individuos, hogares u otro agregado social (Bollen et al. (2001)). El ingreso del hogar constituye una variable clave para representar el nivel socio-económico de los hogares, y por ello es usada en los enfoques más tradicionales de análisis de la pobreza. Un claro ejemplo lo constituye el enfoque de la *línea de la pobreza* en base a ingresos relevados de una encuesta de hogares, usado por muchos países para inferir la situación socioeconómica de la población (Mokomane (2013), Richardson & Bradshaw (2012)).

En muchos casos, la construcción de índices de estatus socioeconómico contempla fines predictivos de diagnóstico para su uso en política focalizadas de detección de pobreza. Un ejemplo de ello lo constituyen aquellos programas focalizados de reducción de la pobreza llevados a cabo por gobiernos y ONGs, en los que se busca clasificar a hogares o individuos entre diferentes grupos socio-económicos a fin de facilitar una ayuda focalizada en aquellos hogares más vulnerables. Este tipo de ayuda suele definirse vía un índice focalizado (comúnmente denominado *Índice de Focalización de Pobreza*) y ha sido utilizado para implementar varios programas de reducción de pobreza (e.g., el CAS en Chile, Sisben en Colombia, SISFOH en Perú, Tekoporá en Paraguay, SIERP en Honduras, y PANES en Uruguay, entre otros). En general, el interés está en la pobreza medida a partir del ingreso (medición directa), por ello dichos programas suelen instrumentarse vía transferencias monetarias o de bienes materiales no asequibles por algunos hogares debido a limitaciones que se derivan de no poder percibir un cierto nivel de ingreso. Por lo tanto, es necesario conocer el monto del ingreso percibido por el hogar al momento de evaluar al potencial beneficiario. Sin embargo, al basarse en un auto-reporte, el ingreso declarado por el potencial beneficiario muy probablemente esté sesgado debido a los incentivos que existen en torno a la ayuda económica y su relación con lo declarado (Doocy & Burnham 2006). Por esta razón, comúnmente se elabora un índice de estatus socio-económico que busca ser una *proxy* o predictor del ingreso, y está basado en variables que resultan más fáciles de observar y captar, tales como los activos del hogar (TV, radio, transporte, etc.), las condiciones habitacionales (e.g. materiales de los techos, pisos y paredes de la vivienda) y otras variables que caracterizan socialmente a los miembros, como su escolaridad y ocupación. La técnica más utilizada para la elaboración de este tipo de índice *SES* es la de Componentes Principales (PCA) (Merola & Baulch (2014), Hoque (2014)), proponiéndose recientemente algunas extensiones para variables categóricas ordinales (Kolenikov & Angeles (2009)).

A pesar del uso extensivo de PCA en la aplicaciones mencionadas, tal método no explota toda la información contenida en las datos *training*, sea para predecir una variable de interés o bien para estimar el efecto marginal del nivel socio-económico sobre una cierta respuesta que se busca explicar. Más precisamente, dicho método no contempla la existencia de alguna variable respuesta de interés Y (por ej. ingreso, pobreza monetaria, fertilidad, consumo, etc.), perdiendo así información relevante para los fines

predictivos. Es decir que si desea predecir el ingreso de un hogar (i.e. ingreso como variable respuesta) utilizando como predictoras un conjunto de variables *proxy* observables y se cuenta con una muestra o sub-muestra de hogares con sus respectivos ingresos monetarios, el índice SES derivado de la reducción vía PCA puede utilizarse para predecir el ingreso de un hogar fuera de dicha muestra pero no utilizaría la información del ingreso relevado por los hogares contenidos en la misma. El enfoque de Reducción Suficiente de Dimensiones (SDR), al igual que PCA, busca reducir el espacio de co-variables o predictoras $\mathbf{X}$ pero a diferencia de PCA utiliza información de la variable respuesta que se está modelando. Específicamente, para un vector de p co-variables $\mathbf{X} \in \mathbb{R}^p$ y una variable respuesta Y, SDR busca una reducción $R(\mathbf{X}) \in \mathbb{R}^d$, $d \leq p$, de forma tal que $Y|\mathbf{X} =_d Y|R(\mathbf{X})$, donde $Y|\mathbf{X}$ denota la distribución condicional de Y dado $\mathbf{X}$, y '$=_d$' denota la igualdad en distribución. Gran parte de la literatura metodológica asociada a SDR está basada en el enfoque de *regresión inversa* (i.e. $\mathbf{X}|Y$) desarrollado por Cook y varios co-autores (e.g. Cook (1998*a*), Cook & Weisberg (1991), Cook (1994, 1998*b*, 2007), Cook & Lee (1999), Bura & Cook (2001), Cook & Yin (2001), Chiaromonte et al. (2002), Cook & Ni (2005), Cook & Forzani (2008, 2009))..

Los métodos de estimación de la reducción $R(\mathbf{X})$ en el marco de SDR pueden clasificarse básicamente en aquellos basados en los momentos de la distribución condicional de $\mathbf{X}|Y$ (e.g. Li (1991), Cook & Weisberg (1991), Li (1992), Bura & Cook (2001), Xia et al. (2002), Li et al. (2005), Cook & Ni (2005), Zhu & Zeng (2006), Cook & Li (2002), Li & Wang (2007)) y aquellos basados en modelos para la regresión inversa $\mathbf{X}|Y$ (e.g. Cook (2007), Cook & Forzani (2008, 2009)). Estas metodologías han sido desarrolladas originalmente en problemas de regresión que involucran predictores continuos. Sin embargo, dadas las características de las variables comúnmente disponibles en las bases de datos sociales, resulta necesario extender la metodología de SDR a otros tipos de variables, en particular a variables categóricas.

Una extensión que puede aplicarse para variables dicotómicas o multinomiales (tipo Bernoulli) ha sido recientemente desarrollada en el marco de *modelos lineales generalizados* por Bura et al. (2016). A pesar de ello, gran parte de las variables usadas para la construcción de índices SES tienen una naturaleza ordinal, no pudiendo encuadrar su distribución dentro de la familia exponencial para la aplicación de tal extensión. Por tal motivo, Forzani et al. (2018) proponen un método de reducción supervisada, basado en el enfoque de RSD en un marco de regresión inversa. Puntualmente, asumiendo un modelo para variables latentes, adoptan el método de componentes principales ajustados (PFC) derivado por Cook & Forzani (2009) para extenderlo a preditores ordinales (PFCORDINAL). En el presente trabajo, vamos a mostrar las ventajas en términos predictivos, de utilizar PFCORDINAL para la construcción de índices SES. La aplicación se realiza utilizando datos de la Encuesta Permanente de Hogares (EPH) de Argentina elaborada por el INDEC.

2. Metodología

2.1. Reducción en un Modelo de Variables Latentes

El problema aplicado de interés podemos representarlo por medio de un modelo de regresión en el cual tenemos una variable respuesta $Y \in \mathbb{R}$ (e.g. ingreso o pobreza) sobre un conjunto de predictores $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$, donde cada X_j, $j = 1, \ldots, p$ es una variable categórica ordenada, es decir $X_j \in \{1, 2, \ldots, G_j\}$, $j = 1, \ldots, p$. Para encontrar una índice *SES* unidimensional a partir de las variables $\mathbf{X}$, esto es una reducción de $\mathbf{X}$, adoptamos el enfoque de regresión inversa basado en modelos (Cook 2007). Para ello, dado el vector de p variables ordinales observadas $\mathbf{X}$, supondremos la existencia de un vector p-dimensional de variables latentes continuas no observadas subyacentes a cada variable ordinal, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$, tal que satisfacen el siguiente modelo

$$\mathbf{Z}|Y = \boldsymbol{\mu}_Y + \boldsymbol{\epsilon}, \tag{1}$$

donde $\boldsymbol{\mu}_Y = \mathrm{E}(\mathbf{Z}|Y)$ y el término de error $\boldsymbol{\epsilon}$ es independiente de Y, normalmente distribuido y con media $\mathbf{0}$ y matriz de covarianza $\boldsymbol{\Delta}$ definida positiva. Como es usual en los modelos de variables latentes para datos ordinales, debemos imponer algunas restricciones para identificar los parámetros del modelo, que en el presente caso serán: $[\boldsymbol{\Delta}]_{jj} \doteq \delta_j = 1$ y $\mathrm{E}(\mathbf{Z}) = 0$ (ver Jackman 2009). En el presente contexto, cada variable observada X_j es una versión discretizada a partir de una variable latente Z_j de la siguiente manera: Para un conjunto de umbrales $\boldsymbol{\Theta}^{(j)} = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{G_j}^{(j)}\}$, dividimos la recta real en intervalos disjuntos $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \ldots < \theta_{G_{j-1}}^{(j)} < \theta_{G_j}^{(j)} = +\infty$ y luego tomamos

$$X_j = \sum_{g=1}^{G_j} g\mathbb{I}(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)}),$$

donde $\mathbb{I}(A)$ es una función indicadora del conjunto A. Esto es, X_j es una variable escalonada derivada de dicretizar Z_j.
En el desarrollo que sigue, vamos a denotar $\boldsymbol{\Theta} \doteq \{\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(p)}\} = \{\theta_0^{(1)}, \ldots, \theta_{G_1}^{(1)}, \ldots\ldots, \theta_0^{(p)}, \ldots, \theta_{G_p}^{(p)}\}$ y $C(\mathbf{X}, \boldsymbol{\Theta}) = [\theta_{X_1-1}^{(1)}, \theta_{X_1}^{(1)}) \times \cdots \times [\theta_{X_p-1}^{(p)}, \theta_{X_p}^{(p)})$. Sea $\mathcal{Y}$ el espacio dominio de Y y $\mathcal{S}_{\boldsymbol{\Gamma}} = \text{span}\{\boldsymbol{\mu}_Y - E(\boldsymbol{\mu}_Y)|Y \in \mathcal{Y}\}$. Si $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, con $d \leq p$, es una matriz semi-ortogonal, cuyas columnas forman una base para el subespacio d-dimensional $\mathcal{S}_{\boldsymbol{\Gamma}}$, siguiendo a Cook & Forzani (2008), podemos re-escribir el modelo (1) de la forma

$$\mathbf{Z}|Y = \boldsymbol{\Gamma}\boldsymbol{\nu}_Y + \boldsymbol{\epsilon}, \tag{2}$$

donde $\boldsymbol{\nu}_Y = \boldsymbol{\Gamma}^T\boldsymbol{\mu}_Y \in \mathbb{R}^d$ con $E(\boldsymbol{\nu}_Y) = 0$ y $\mathrm{var}(\boldsymbol{\nu}_Y) > 0$. En este caso modelamos al vector de coordenadas de la forma $\boldsymbol{\nu}_Y = \boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\}$ donde $\mathbf{f}_Y \in \mathbb{R}^r$ es un vector de r funciones conocidas de Y tales que $E((\mathbf{f}_Y - E(\mathbf{f}_Y)(\mathbf{f}_Y - E(\mathbf{f}_Y))^T)$ conforma una matriz definida positiva y $\boldsymbol{\xi} \in \mathbb{R}^{d\times r}$ es una matriz de rango completo d, con $d \leq r$ (ver Cook & Forzani 2008). Bajo este modelo, cada coordenada de $\mathbf{Z}|Y$ es modelada linealmente como función de un vector de predictores dado por $\mathbf{f}_Y$, y por lo tanto, cuando Y es cuantitativa, podemos usar gráficas inversas con el fin de obtener información para seleccionar la función $\mathbf{f}_Y$, lo que no es posible en la regresión original de Y sobre $\mathbf{X}$ debido a sus dimensiones. Cuando Y es continua, $\mathbf{f}_Y$ usualmente quedará representada por un conjunto flexible de funciones básicas, como ser un polinomio en Y, lo cual resulta ser una opción parsimoniosa cuando el método gráfico se ve agotado o resulta poco práctico para todos los predictores. Cuando Y es categórica y toma valores tales como $\{C_1, \ldots, C_h\}$, podemos tomar $r = h - 1$ y especificar j-ésimo elemento de $\mathbf{f}_Y$ a través de $\mathbb{I}(y \in C_j)$, con $j = 1, \ldots, h$. Cuando Y es continua también podemos particionar sus valores en h categorías $\{C_1, \ldots, C_h\}$ y luego especificar para la coordenada j-ésima de $\mathbf{f}_Y$ de la misma manera que para el caso en que Y sea categórica. De esta manera, el modelo (2) puede expresarse como

$$\mathbf{Z}|Y = \boldsymbol{\Gamma}\boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\} + \boldsymbol{\epsilon}, \tag{3}$$

donde $\boldsymbol{\epsilon}$ es independiente de Y, normalmente distribuido con media $\mathbf{0}$ y matriz de varianza-covarianza (definida positiva) $\boldsymbol{\Delta}$, con unos en la diagonal a los fines de la identificabilidad del modelo. Si $\mathbf{Z}$ fuera observable, luego bajo el modelo (3) la reducción suficiente minimal vendría dada por $R(\mathbf{Z}) = \boldsymbol{\alpha}^T\mathbf{Z}$, con $\boldsymbol{\alpha} \equiv \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ en base al Teorema 2.1 de Cook & Forzani (2008). A su vez debe notarse que si $R(\mathbf{Z}) = \alpha^T\mathbf{Z}$ es una reducción suficiente para $Y|\mathbf{Z}$, luego $\tilde{R}(\mathbf{Z}) = A\alpha^T\mathbf{Z}$ también es una reducción suficiente para dicha regresión, para cualquier matriz A invertible de orden $d \times d$. Por ende, lo que resulta identificable es el $\mathrm{span}(\alpha)$ y no α en sí mismo.

Volviendo al problema original, lo que se busca para construir el índice SES es una reducción para la regresión $Y|\mathbf{X}$ (i.e. para los predictores ordinales originales). Como bien notan Forzani et al. (2018) ,$\mathbf{X}$ es una función de $\mathbf{Z}$, por lo que $\boldsymbol{\alpha}^T\mathbf{Z}$ será una reducción suficiente para $Y|\mathbf{X}$, i.e. $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{Z}$ (Cook (1998*a*)). Sin embargo, $\mathbf{Z}$ no es observada y la única información disponible es $\mathbf{X}$. Sobre esta base, Forzani et al. (2018) proponen tomar la esperanza condicional de $\boldsymbol{\alpha}^T\mathbf{Z}$ dado $\mathbf{X}$, en lugar de $\boldsymbol{\alpha}^T\mathbf{Z}$ para la reducción de $Y|\mathbf{X}$, puesto que constituye el mejor predictor de $\boldsymbol{\alpha}^T\mathbf{Z}$ en términos del minimo error

cuadrático medio. Por ende, la reducción supervisada para $Y|\mathbf{X}$ queda definida por

$$R(\mathbf{X}) = \mathrm{E}(\boldsymbol{\alpha}^T\mathbf{Z}|\mathbf{X}), \tag{4}$$

y tomando $d = 1$, el índice de estatus socioeconómico será $SES(\mathbf{X}) \doteq R(\mathbf{X}) = \boldsymbol{\alpha}^T\mathrm{E}(\mathbf{Z}|\mathbf{X}) \in \mathbb{R}$. Contrariamente a los índices tipo PCA, este nuevo enfoque usa la información acerca de la respuesta bajo análisis.

2.2. Estimación del índice *SES*

De lo anterior surge que para obtener el índice es necesario estimar los parámetros de la reducción $\boldsymbol{\alpha} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$. Si $\mathbf{Z}$ fuera observada, el estimador de máxima versoimilitud de $\boldsymbol{\alpha}$ estaría dado por $\widehat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}\mathbf{V}_d$, donde $\mathbf{V}_d$ son los primeros d autovectores de $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}\hat{\boldsymbol{\Sigma}}_{\mathbf{res}}^{-1/2}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}$ es la matriz de covarianza muestral de los vectores ajustados de la regresión lineal multivariada de $\mathbf{Z}$ sobre $\mathbf{f}_Y$ y $\hat{\boldsymbol{\Sigma}}_{\mathbf{res}} = \hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\mathbf{fit}}$, siendo $\hat{\boldsymbol{\Sigma}}$ la matriz de covarianza muestral (marginal) de los predictores; esto es, el método de Componentes Principales Ajustadas (PFC). Sin embargo, $\mathbf{Z}$ es una variable latente no observada, y por lo tanto, las matrices de covarianza (marginal y ajustada) no pueden observarse de forma directa. En vista de la robustez probada en Cook & Forzani (2008), podríamos considerar la aplicación de la metodología de PFC directamente sobre $\mathbf{X}$ de manera *naive*, y aún así obtendríamos un estimador $\sqrt{n}$-consistente. Este enfoque constituirá la base para comparar el método para ordinales. También será usado para obtener los valores iniciales del algoritmo propuesto para obtener estimadores de máxima verosimilitud bajo el modelo de variables latentes.

Para estimar los parámetros del modelo (3) y en particular los parámetros de la reducción usaremos la parametrización $\boldsymbol{\alpha} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ de forma tal que (3) puede ser re-expresado de la forma

$$\begin{aligned} \mathbf{Z}|Y &= \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\{\mathbf{f}_Y - E(\mathbf{f}_Y)\} + \boldsymbol{\epsilon}, \\ \text{con } \ \boldsymbol{\alpha}^T\boldsymbol{\alpha} &= \mathbf{I} \ \text{ y } \ \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Delta}), \\ \text{y } \ X_j = g &\Leftrightarrow Z_j \in [\theta_{g-1}^{(j)}, \theta_g^{(j)}),\ j = 1, \ldots, p. \end{aligned} \tag{5}$$

Supongamos que tenemos una muestra aleatoria de n puntos $(y_i, \mathbf{x}_i)$ extraídos de la distribución conjunta de $(Y, \mathbf{X})$ tal que satisfacen el modelo (5). Necesitamos estimar el span$(\boldsymbol{\alpha})$. Supongamos por el momento que la dimensión d de la reducción es conocida (de hecho, para la construcción del índice SES tomamos $d = 1$). Para obtener el estimador con los datos observados, necesitamos maximizar la función de log-verosimilitud

$$\sum_{i=1}^{n} \log f_{\mathbf{X}}(\mathbf{x}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}). \tag{6}$$

Sea $C(\mathbf{X}, \boldsymbol{\Theta})$ el hiper-cubo $C(\mathbf{X}, \boldsymbol{\Theta}) = [\theta_{X_1-1}^{(1)}, \theta_{X_1}^{(1)}) \times \cdots \times [\theta_{X_p-1}^{(p)}, \theta_{X_p}^{(p)})$. Como para cada $g = 1, \ldots, G_j$, $X_j = g \Leftrightarrow Z_j \in [\theta_{X_j-1}^{(j)}, \theta_{X_j}^{(j)})$ y $\mathbf{Z}|Y$ está distribuida normalmente, para cada i, la densidad truncada (no normalizada[1]) $f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi})$ es

$$f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = (2\pi)^{-p/2}|\boldsymbol{\Delta}|^{-1/2}e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Delta}^{-1}(\mathbf{z}_i - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})(\mathbf{z}_i - \boldsymbol{\Delta}\boldsymbol{\alpha}\boldsymbol{\xi}\bar{\mathbf{f}}_{y_i})^T)}I_{\{\mathbf{z}_i \in C(\mathbf{x}_i, \boldsymbol{\Theta})\}},$$

donde $\bar{\mathbf{f}}_{y_i} \doteq \mathbf{f}_{y_i} - n^{-1}\sum_{i=1}^{n}\mathbf{f}_{y_i}$. Por lo tanto, para cada i, la densidad marginal no normalizada de $\mathbf{X}|Y$ será

$$f_{\mathbf{X}}(\mathbf{x}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi}) = \int f_{\mathbf{X},\mathbf{Z}}(\mathbf{x}_i, \mathbf{z}_i|y_i; \boldsymbol{\Theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\xi})\, d\mathbf{z}_i.$$

Debido a que el cómputo exacto de la función de verosimilitud resulta muy complejo debido a las integrales

[1] El término 'no normalizada' se utiliza, debido a que en un sentido estricto, no constituye una densidad al no integrar 1. No obstante, la literatura relacionada hace caso omiso de ello en el sentido de que sigue denominándola densidad.

múltiples contenidas en la misma, en estos casos, los estimadores de máxima verosimilitud comúnmente se obtienen usando un algoritmo iterativo EM (*Expectation-Maximization*). Este alternativa es usual en modelos con variables latentes, debido a que reduce la complejidad de cómputo de la verosimilitud conjunta de $(\mathbf{X}, \mathbf{Z})$. Para detalles del algoritmo EM para la estimación de este modelo, puede verse Forzani et al. (2018).

Por otra parte, cuando realizamos reducción suficiente de dimensiones para la construcción del *SES* a partir de una combinación lineal de de los predictores originales (tomando como dimensión de la reducción $d = 1$), tal combinación lineal típicamente involucra a todas las variables originales. Esto significa que aún las variables no relevantes o redundantes son incluidas en el cómputo final, dificultando más la interpretación. Para superar esta limitación, se puede realiza selección de variables y de este modo obtener combinaciones lineales que sólo incluyan a las variables activas o relevantes. Puntualmente, podemos inducir selección de variables en reducción de dimensiones, agregando una penalización en el proceso de maximización inmerso en el algoritmo EM. Esta penalización es del tipo *group-lasso* ya que, con el fin de no elegir una variable X_j en particular, necesitamos hacer que toda la fila j-ésima de $\boldsymbol{\alpha}$ se iguale a 0. Por ello, siguiendo a Chen et al. (2010), utilizamos una norma mixta tipo ℓ_1/ℓ_2, donde la norma interna es la norma ℓ_2 de cada fila de $\boldsymbol{\alpha}$. El término de parámetro de penalización λ puede seleccionarse utilizando algún criterio de información, como ser el criterio de Akaike (AIC) o el de Bayes (BIC). Los detalles pueden encontrarse en Chen et al. (2010). Otra alternativa es encontrar el valor de λ que minimiza el error de predicción vía un experimento de validación cruzada, pero esto requiere la adopción de alguna regla de predicción.

Cabe destacar que este procedimiento seguido por Forzani et al. (2018) realiza al mismo tiempo selección de variables y reducción suficiente de dimensiones sin necesidad de especificar un modelo para $Y|\mathbf{X}$. Entonces, la reducción obtenida puede utilizarse luego con cualquier regla de predicción. Este enfoque difiere de aquellos en los que el procedimiento de selección de variables hace uso de un modelo de regresión en particular, como por ejemplo en Gertheiss & Tutz (2010).

2.3. Datos y Variables

Los datos utilizados provienen de la base de microdatos de la *Encuesta Permanente de Hogares* (EPH) de Argentina, tomando, en particular, el cuarto trimestre de 2013. La EPH es la principal encuesta de hogares de Argentina y es realizada por el *Instituto Nacional de Estadísticas y Censos* (INDEC).
Consideramos nueve variables ordinales sobre las condiciones de vida de los hogares, y dos variables socio-económicas sobre el jefe/a de hogar (nivel de instrucción formal y situación laboral). Específicamente, las variables predictoras utilizadas para la construcción de los índices *SES* son:

- *Ubicación de la vivienda*: Indica si la vivienda está ubicada en una zona desfavorable o en una área vulnerable. Más precisamente, esta variable indica si la vivienda: (i) está ubicada en una zona inundable, (ii) o/y cerca de un basural, (iii) o/y en una villa de emergencia. Esta variable tiene 4 categorías: vale 1 si la vivienda presenta conjuntamente las características (i)-(iii), 2 para viviendas que presentan dos de las características (i)-(iii), 3 si la vivienda posee solo una de estas características, y 4 si la vivienda no tienen ninguna de tales características.

- *Calidad de la vivienda*: Contempla de forma conjunta la calidad (materiales) del techo, paredes y pisos de la vivienda, en base a la la metodología CALMAT usada para el Censo Poblacional de Argentina. Tiene 4 categorías en orden creciente en términos de la calidad de la vivienda (i.e. asume un valor mayor, cuando la calidad de la vivienda es superior).

- *Combustible para cocinar*: Indica el tipo de combustible predominante en la vivienda para la preparación de los alimentos. Tiene 3 categorías: vale 1 si el principal combustible para cocinar es kerosene, madera o carbón, 2 si tiene gas envasado, y 3 si posee gas natural por tubería.

- *Hacinamiento*: Caracteriza el hacinamiento del hogar y se deriva del cómputo del ratio entre ambientes de la vivienda y la cantidad de miembros del hogar. Tiene 4 categorías: 1 si este ratio es menor o igual a 1, 2 si el ratio está en el intervalo (1,2], 3 si este está en (2,3], y 4 el ratio es mayor a 3.
- *Escolaridad*: Indica el nivel de instrucción formal alcanzado por el jefe/a del hogar. Tiene 7 categorías: 1 si el jefe/a de hogar no posee educación formal, 2 si tiene primaria incompleta, 3 si tiene primaria completa, 4 si realizó secundaria incompleta, 5 si realizó escuela secundaria completa, 6 si tiene estudios superiores incompletos y 7 si el jefe/a posee titulo terciario o universitario.
- *Horas trabajadas*: Describe las situación laboral del jefe/a de hogar. Tiene 4 categorías: 1 si está desempleado o inactivo, 2 cuando el jefe/a de hogar trabaja menos de 40 horas semanales, 3 para 40-45 horas semanales de trabajo, y 4 cuando el jefe/a de hogar está empleado con más de 45 horas semanales.
- *Desagüe*: Indica el tipo de drenaje o desagüe que posee el baño la vivienda. Tiene 4 categorías: 1 si el desagüe consisten en un agujero, 2 si el desagüe es en un pozo negro, 3 si es pozo negro con cámara séptica, y 4 para tuberías de desagüe en una red pública.
- *Instalación sanitaria*: Indica el tipo de instalación sanitaria que posee la vivienda. Tiene 3 categorias: 1 para letrinas, 2 para baño con inodoro sin descarga de agua, y 3 para baños con descarga de agua.
- *Baño compartido*: Indica si el baño es compartido o no. Tiene 3 categorías: 1 si el baño está fuera de la vivienda y es compartido con otras, 2 si el baño es compartido con otros hogares dentro de la vivienda, y 3 si el baño es de uso exclusivo del hogar.
- *Ubicación del agua*: Indica la ubicación más cercana para obtener agua potable. Tiene 3 categorías: 1 si el agua potable se consigue fuera del terreno de la vivienda, 2 si el agua se obtiene dentro del terreno, pero fuera de la vivienda, y 3 si el agua potable se obtiene en el interior vivienda por tubería.
- *Provisión de agua*: Indica la fuente de donde proviene el agua de la vivienda. Tiene 3 categorías: 1 si el agua potable proviene de una bomba de mano o de un grifo público compartido con los vecinos, 2 si el agua potable se obtiene mediante una bomba de perforación automatizada, y 3 si la vivienda tiene agua potable por tubería.

A fin de tener en cuenta la heterogeneidad regional, estimamos índices SES de forma separada para cada una de las siguientes cinco regiones: región metropolitana del Gran Buenos Aires ($n = 2351$ hogares), región Pampeana ($n = 5003$), el Noroeste Argentino (NOA) ($n = 2852$), el Noreste Argentino (NEA) ($n = 1594$), y la Patagonia ($n = 2398$).

Para evaluar el nivel predictivo, consideramos dos tipos de respuesta: una continua, dada por el ingreso per cápita familiar (`ipcf`), y una binaria basada en el ingreso (`pobreza`) que indica si el hogar está por encima o debajo de la linea de pobreza (i.e. si el hogar es pobre o no, en términos de ingresos).

3. Resultados

Estamos interesados en mostrar que el método de reducción supervisada porpuesto por Forzani et al. (2018), que denominaremos REG-PFCORD, constituye una alternativa superior a la metodologías tipo PCA para la construcción de índices SES, al mismo tiempo que logra tener un poder predictivo similar a considerar el conjunto entero de predictores (i.e. sin aplicar reducción). Para ello, el desempeño en predicción del índice propuesto (REG-PFCORD) se compara con las siguientes estrategias metodológicas:

- Consideración de todo el conjunto de predictores sin aplicar reducción, tratados como predictores continuos, en el sentido métrico. Vamos a denominar a este método FULL.

- Todo el conjunto de predictores incluidos sin aplicar reducción, incorporados por medio de variables *dummies*. Vamos a llamar a este enfoque FULL-I.
- Todo el conjunto de predictores es incluido pero usando un procedimiento tipo *group-lasso* para predictores ordinales (Gertheiss & Tutz 2010), realizando con ello selección de variables. A este método lo llamaremos LASSOORD.
- Una variante del PCA diseñada para predictores ordinales usando correlaciones policóricas (Kolenikov & Angeles 2009). Lo denominaremos PCAPOLY.
- Una variante no lineal del PCA que utiliza un escalamiento especial para aplicarse a categorías ordinales (Linting & van der Kooij 2009). A este método lo llamaremos NLPCA.

Las primeras tres estrategias son incluidas con el objetivo de tener una base de referencia sobre el desempeño que se obtiene cuando se usa el conjunto entero de predictores, pero debe quedar claro que las mismas no brindan un índice. De hecho, solamente las dos últimas alternativas de la lista compiten en la construcción del índice SES con la metodología REG-PFCORD.

Para cada estrategia, ajustamos una regresión logística para la respuesta `pobreza` (discreta) y una regresión lineal para la variable respuesta `ipcf` (continua).Cuando computamos la reducción suficiente, elegimos una $\mathbf{f}_Y$ diferente para cada respuesta. Para la respuesta continua, usamos una base polinómica de grado $r = 2$. Para la respuesta binaria, $\mathbf{f}_Y$ es simplemente una variable indicadora centrada.

Los datos son particionados en diez conjuntos disjuntos a fin de tener diez réplicas experimentales. En cada corrida, uno de los subconjuntos es usado como conjunto de prueba, mientras que el resto de ellos conforman la muestra de entrenamiento. Con cada método se obtiene el error cuadrático medio (MSE) de realizar validación cruzada con diez iteraciones (10-fold cross-validation). Los resultados se muestran en el cuadro 1, reportándose también los correspondientes desvíos estandar.

En primer lugar, de la tabla puede observarse que, para la respuesta continua, usar variables dummy para el conjunto entero de predictores (FULL-I) o realizar selección de variables con LASSOORD, constituyen estrategias más efectivas que considerar el conjunto entero de predictores como variables continuas (FULL). Lo contrario ocurre cuando consideramos la respuesta binaria; esto es FULL en general brinda mejores resultados predictivos que FULL-I y LASSOORD.

Para los índices SES, tanto para la respuesta continua como discreta, los resultados muestran que REG-PFCORD es superior que PCAPOLY y NLPCA. A su vez, entre los índices tipo PCA, NLPCA muestra en general un mejor desempeño que PCAPOLY, excepto para algunos casos donde la respuesta es discreta. Por otra parte, los errores de predicción obtenidos con el índice REG-PFCORD son muy similares a aquellos que surgen de la predicción con FULL. Cuando consideramos la respuesta discreta, el REG-PFCORD da mejores resultados predictivos que el LASSOORD para todas las regiones consideradas. También debemos remarcar que, contrariamente al LASSOORD, los índices obtenidos usando la técnica de reducción suficiente de dimensiones, nos permite utilizarlos con cualquier método predictivo.

Para ilustrar mejor el ajuste obtenido, la Figura 1 muestra gráficamente el resultado del modelo de regresión de `ipcf` sobre el índice SES obtenido usando todos los datos. Un término cuadrático SES^2 es incluido para corregir la curvatura en la función de regresión estimada y la variable respuesta fue transformada con $\text{ipcf} \leftarrow \text{ipcf}^{1/3}$, a partir de realizar un análisis de regresión Box–Cox. Puede observarse que cuando usamos el índice SES construido con el PCAPOLY, los valores del índice están concentrados principalmente en el intervalo [0.5, 1.0] mientras que, para el SES modelado usando REG-PFCORD, los valores del indice se distribuyen más regularmente sobre todo el intervalo [0, 1.0]. Esto permite obtener un mejor ajuste con el modelo lineal, como se revela a partir un valor del R^2 igual a 0.302 comparado con un 0.231 obtenido con un índice SES basado en PCA.

Cuadro 1: MSE para el índice SES (10-fold cross-validation)

		Errores de Predicción -MSE				
Respuesta	Método	*Buenos Aires*	*Pampa Húmeda*	*NEA*	*NOA*	*Patagonia*
Ingreso per capita	REG-PFCORD	7.29	4.72	4.73	3.34	12.80
(continua)		(2.91)	(1.73)	(2.69)	(1.52)	(3.72)
	PCApoly	7.60	5.10	5.07	3.68	14.7
		(2.45)	(0.90)	(1.77)	(0.90)	(4.01)
	NLPCA	7.38	4.95	4.89	3.52	13.67
		(2.29)	(0.61)	(1.48)	(0.65)	(3.71)
	FULL	7.22	4.69	4.68	3.32	13.14
		(3.50)	(0.88)	(2.47)	(0.93)	(3.34)
	FULL-I	7.01	4.52	4.48	3.08	12.92
		(2.46)	(0.83)	(1.74)	(0.76)	(3.80)
	LASSOord	7.00	4.51	4.42	3.05	12.88
		(2.46)	(0.83)	(1.77)	(0.76)	(3.84)
Pobreza	reg-PFCord	0.204	0.169	0.278	0.288	0.126
(discreta)		(0.017)	(0.012)	(0.031)	(0.029)	(0.025)
	PCApoly	0.213	0.188	0.324	0.357	0.133
		(0.024)	(0.020)	(0.026)	(0.053)	(0.027)
	NLPCA	0.212	0.188	0.325	0.358	0.134
		(0.023)	(0.019)	(0.025)	(0.055)	(0.027)
	FULL	0.202	0.162	0.274	0.287	0.129
		(0.021)	(0.008)	(0.026)	(0.036)	(0.020)
	FULL-I	0.206	0.171	0.286	0.298	0.126
		(0.023)	(0.021)	(0.024)	(0.065)	(0.028)
	LASSOord	0.206	0.171	0.286	0.298	0.129
		(0.023)	(0.021)	(0.024)	(0.065)	(0.027)

Nota: desvíos estándar entre paréntesis. Database: EPH (2013)

Los cuadros 2 y 3 muestran los vectores de coeficientes estimados que definen el índice SES usando `ipcf` y `pobreza` como variable respuesta, respectivamente. Podemos notar que para el método propuesto, varios coeficientes del $\widehat{\alpha}$ han sido llevados a cero con la estimación regularizada, mientras que para PCAPOLY y NLPCA sólo *horas trabajadas* pareciera no ser muy relevante para la construcción del índice SES. Además, de los resultados reportados podemos apreciar varias diferencias entre el REG-PFCORD y los enfoques basados en PCA (i.e., PCAPOLY y NLPCA).

En primer lugar, la importancia de cada predictor para la construcción de índice SES difiere según el método. Por ejemplo, la variable *hacinamiento* obtiene la mayor ponderación con el REG-PFCORD para todas las regiones y para ambas variables respuesta, mientras que la *instalación sanitaria* y la *ubicación del agua* muestran ser las más relevantes para la construcción del índice usando PCAPOLY, y la *instalación sanitaria* y el *desagüe* para el método NLPCA.

En segundo lugar, observamos que los índices SES construidos usando ambos métodos basados en PCA dan similares ponderaciones a los predictores para todas las regiones. En cambio, el índice SES basado en REG-PFCORD logra captura la divergencia económica regional explicada por diferencias en dotaciones de factores, productividad, niveles de actividad y patrones de crecimiento económico regional. Más aún, en las regiones urbanas más ricas (específicamente, Buenos Aires y la región Pampeana) la estimación regularizada del REG-PFCORD tiende a fijar en cero las variables con mayor ponderación en los índices basados en PCA. Esta diferencia resulta de especial interés, ya que estas regiones suelen tener una mejor infraestructura social básica general, por lo que las variables relacionadas con el drenaje, provisión de agua e la instalación sanitaria de la vivienda son menos importantes para medir el nivel socioeconómico, pues gran parte de la población estaría cubierta en este sentido. Con un fundamento similar encontramos

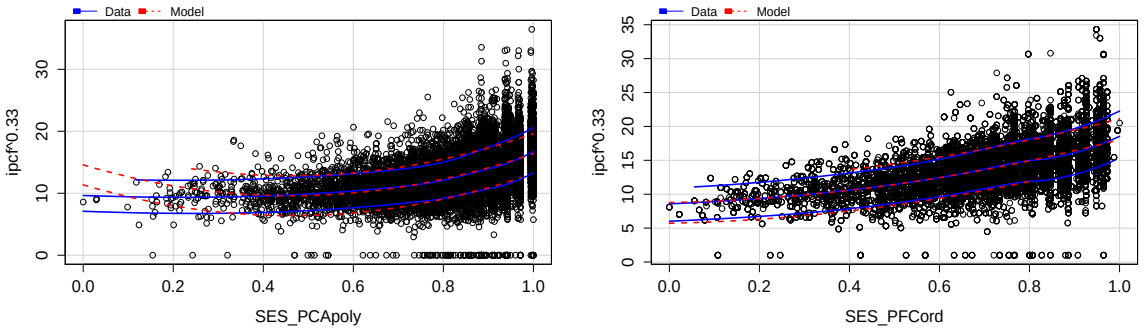

Figura 1: Ajuste del modelo lineal del ingreso per cápita como función del índice SES. Lado izquierdo: usando índice SES con método tipo PCA (PCAPOLY). Lado derecho: resultado con el método REG-PFCORD.

que otras variables, tal como *hacinamiento* o *escolaridad*, son necesarias en orden de obtener un mejor índice SES para predecir el nivel de ingreso del hogar. En la misma línea, para regiones con mayores niveles de pobreza (NOA y NEA) el índice SES provisto por el método REG-PFCORD muestra que otras variables, tal como *ubicación de la vivienda*, *provisión de agua* o *ubicación del agua*, pasan a ser relevantes en la determinación del nivel socio-económico.

Comparando ambas metodologías basadas en PCA, podemos notar que el NLPCA es más sensible a la heterogeneidad regional que el PCAPOLY, pero las diferencias en las ponderaciones del índice de estos métodos comparado con las del REG-PFCORD permanecen sustanciales.

Adicionalmente, podemos apreciar que el índice SES obtenido vía REG-PFCORD es sensible respecto a qué variable respuesta estamos usando para caracterizar el fenómeno social de interés. Por ejemplo, en las regiones del GBA, Pampeana y Patagonia, la *escolaridad* tiene un peso considerable en el índice SES para explicar el ingreso per cápita, no así para predecir pobreza. Esto cobra sentido debido a que, en estas regiones relativamente más ricas, el acceso a niveles básicos de escolaridad está más garantizado en la población que habita en las mismas. Por otra parte, muchas veces el ingreso constituye un fuerte incentivo sobre la decisión individual de alcanzar mayores niveles de educación, reflejándose de hecho en la ponderación del índice para predecir ingreso. Además, para estas regiones más ricas, algunas variables, tal como *desagüe*, *instalación sanitaria* o *baño compartido* pasan a ser relevantes para predecir si el hogar es pobre o no (siguiendo el criterio de la línea de pobreza). Tales diferencias no pueden ser capturas por el índice SES basado en la metodología de PCA.

Cuadro 2: Comparación de coeficientes del índice SES para las metodologías REG-PCFORD, PCAPOLY y NLPCA para predecir ingreso per cápita del hogar.

Variables	Buenos Aires			Pampa Húmeda			NOA		
	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA
Ubicación de la vivienda	0	-0.1690	-0.0943	0	-0.1903	-0.0976	-0.1314	-0.1068	-0.0835
Calidad de la vivienda	-0.1985	-0.3768	-0.2199	0.2591	-0.3557	-0.1985	0	-0.3278	-0.1849
Combustible para cocinar	-0.4646	-0.3788	-0.2080	0.3627	-0.3609	-0.1678	-0.1070	-0.3287	-0.1582
Hacinamiento	-0.7272	-0.2888	-0.1788	0.8300	-0.2351	-0.1329	-0.8798	-0.1991	-0.1194
Escolaridad	-0.2676	-0.2275	-0.1474	0.2668	-0.2075	-0.1135	-0.3614	-0.2197	-0.1201
Desagüe	-0.0873	-0.3381	-0.2047	0	-0.3519	-0.2333	-0.0901	-0.3623	-0.2462
Instalación sanitaria	0	-0.4061	-0.2246	0	-0.4105	-0.2411	0	-0.4217	-0.2545
Baño compartido	0	-0.2759	-0.1186	0	-0.3176	-0.1699	0	-0.2579	-0.1334
Ubicación del agua	0.1700	-0.3918	-0.1790	0	-0.3933	-0.1941	-0.2054	-0.4202	-0.2309
Provisión de agua	0	-0.2023	-0.1033	0	-0.2461	-0.1129	0	-0.3646	-0.1374
Horas trabajadas	-0.3283	0	0	0.2029	0	0	-0.1277	0	0
	NEA			Patagonia					
	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA			
Ubicación de la vivienda	-0.1509	-0.1809	-0.0978	-0.1149	-0.1437	-0.0981			
Calidad de la civienda	0	-0.3727	-0.2130	-0.3046	-0.3258	-0.1844			
Combustible para cocinar	-0.0742	-0.1648	-0.0646	-0.1797	-0.4026	-0.1810			
Hacinamiento	-0.8496	-0.2052	-0.1040	-0.7263	-0.2207	-0.0984			
Escolaridad	-0.3507	-0.1869	-0.1009	-0.3670	-0.1284	-0.0516			
Desagüe	0	-0.3572	-0.2573	-0.1383	-0.4122	-0.2566			
Instalación sanitaria	0	-0.4383	-0.2735	0	-0.4376	-0.2622			
Baño compartido	-0.2284	-0.2921	-0.1344	-0.1204	-0.2937	-0.1734			
Ubicación del agua	0	-0.4227	-0.2377	0.1877	-0.4169	-0.2196			
Provisión de agua	-0.2574	-0.3733	-0.1384	0.2473	-0.1525	-0.0522			
Horas trabajadas	-0.0922	0	0	-0.2637	0	0			

Cuadro 3: Comparación de coeficientes del índice SES para las metodologías REG-PCFORD, PCAPOLY y NLPCA para predecir pobreza (respuesta discreta).

Variables	Buenos Aires			Pampa Húmeda			NOA		
	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA
Ubicación de la vivienda	0	-0.1690	-0.0943	0	-0.1903	-0.0976	-0.2434	-0.1068	-0.0835
Calidad de la vivienda	-0.4033	-0.3768	-0.2199	0.3347	-0.3557	-0.1985	0	-0.3278	-0.1849
Combustible para cocinar	-0.5240	-0.3788	-0.2080	0.3579	-0.3609	-0.1678	0	-0.3287	-0.1582
Hacinamiento	-0.7076	-0.2888	-0.1788	0.7216	-0.2351	-0.1329	-0.7939	-0.1991	-0.1194
Escolaridad	0	-0.2275	-0.1474	0	-0.2075	-0.1135	-0.2094	-0.2197	-0.1201
Desagüe	0	-0.3381	-0.2047	0	-0.3519	-0.2333	0	-0.3623	-0.2462
Instalación sanitaria	0	-0.4061	-0.2246	0.3990	-0.4105	-0.2411	-0.1528	-0.4217	-0.2545
Baño compartido	-0.1836	-0.2759	-0.1186	0	-0.3176	-0.1699	0	-0.2579	-0.1334
Ubicación del agua	-0.1208	-0.3918	-0.1790	0.2647	-0.3933	-0.1941	-0.4933	-0.4202	-0.2309
Provisión del agua	0	-0.2023	-0.1033	0	-0.2461	-0.1129	0	-0.3646	-0.1374
Horas trabajadas	-0.1173	0	0	0.0990	0	0	0	0	0
	NEA			Patagonia					
	reg-PFCord	PCApoly	NLPCA	reg-PFCord	PCApoly	NLPCA			
Ubicación de la vivienda	-0.1982	-0.1809	-0.0978	-0.1187	-0.1437	-0.0981			
Calidad de la vivienda	0	-0.3727	-0.2130	-0.3693	-0.3258	-0.1844			
Combustible para cocinar	-0.2509	-0.1648	-0.0646	-0.2788	-0.4026	-0.1810			
Hacinamiento	-0.7063	-0.2052	-0.1040	-0.3987	-0.2207	-0.0984			
Escolaridad	-0.1442	-0.1869	-0.1009	-0.0766	-0.1284	-0.0516			
Desagüe	0	-0.3572	-0.2573	-0.1887	-0.4122	-0.2566			
Instalación sanitaria	0	-0.4383	-0.2735	-0.1313	-0.4376	-0.2622			
Baño compartido	-0.3477	-0.2921	-0.1344	-0.1289	-0.2937	-0.1734			
Ubicación del agua	0	-0.4227	-0.2377	0.2785	-0.4169	-0.2196			
Provisión de agua	-0.5071	-0.3733	-0.1384	0.6585	-0.1525	-0.0522			
Horas trabajadas	0	0	0	-0.1626	0	0			

4. Conclusiones

En esta presentación se muestra el uso de una nueva metodología de reducción supervisada para la construcción de indicadores de estatus socioeconómico, basado en el enfoque de Reducción Suficiente de Dimensiones. En particular, dicha metodología es una extensión para predictores de naturaleza ordinal, lo que resulta ideal para su uso con datos provenientes de encuestas sociales o microdatos, dado que las variables predominantes son del tipo categóricas ordinales. A priori, el enfoque supervisado resulta superador al considerar información de una variable respuesta de interés, lo que es ignorado por métodos de reducción no supervisados.

La aplicación de esta metodología para la construcción de índices *SES* tomando datos de la Encuesta Permanente de Hogares (EPH) de Argentina, mostró muchas ventajas con respecto a los índices basados en metodos no supervisados (en particular, usando extensiones de PCA para variables ordinales). En particular, el método no sólo ayuda a obtener mejores predicciones sino que también permite obtener una comprensión mejor de las relaciones entre los predictores y la respuesta. De manera más precisa, para el índice de SES, el método supervisado brinda diferentes ponderaciones logrando capturar diferencias regionales, históricas y/o culturales, siendo al mismo tiempo sensible respecto a la medida usada como respuesta (como el ingreso familiar per cápita o la línea de pobreza), lo que no ocurre con las metodologías no supervisadas basadas en PCA. Esta propiedad del método de reducción de dimensiones supervisada basada en modelos tiene implicaciones relevantes para el análisis social aplicado.

Referencias

Bollen, K., Glanville, J. & Stecklov, G. (2001), 'Socioeconomic status and class in studies on fertility and health in developing countries', *Ann. Rev. Sociol.* **27**, 153–185.

Bura, E. & Cook, R. D. (2001), 'Estimating the structural dimension of regressions via parametric inverse regression', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(2), pp. 393–410.

Bura, E., Duarte, S. & Forzani, L. (2016), 'Sufficient reductions in regressions with exponential family inverse predictors. preprint', *Journal of the American Statistical Association* **111**(515), 1313–1330.

Chen, X., Zou, C. & Cook, R. D. (2010), 'Coordinate-independent sparse sufficient dimension reduction and variable selection', *Ann. Statist.* **38**(6), 3696–3723.
URL: *http://dx.doi.org/10.1214/10-AOS826*

Chiaromonte, F., Cook, R. & Li, B. (2002), 'Sufficient dimensions reduction in regressions with categorical predictors', *Ann. Statist.* **30**(2), 475–497.

Cook, R. (1994), Using dimension reduction subspaces to identify important inputs in models of physical systems, *in* 'Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association', pp. 18–25.

Cook, R. (1998*a*), *Regression Graphics*, Wiley, New York.

Cook, R. (2007), 'Fisher lecture: Dimension reduction in regression (with discussion)', *Statistical Science* **22**, 1–26.

Cook, R. D. (1998*b*), 'Principal hessian directions revisited', *Journal of the American Statistical Association* **93**(441), 84–94.

Cook, R. D. & Lee, H. (1999), 'Dimension reduction in binary response regression', *Journal of the American Statistical Association* **94**(448), 1187–1200.

Cook, R. D. & Ni, L. (2005), 'Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach', *Journal of the American Statistical Association* **100**(470), 410–428.

Cook, R. & Forzani, L. (2008), 'Principal fitted components for dimension reduction in regression', *Statistical Science* **23**, 485–501.

Cook, R. & Forzani, L. (2009), 'Likelihood-Based sufficient dimension reduction', *Journal of the American Statistical Association* **104**(485), 197–208.

Cook, R. & Li, B. (2002), 'Dimension reduction for conditional mean in regression', *The Annals of Statistics* **30**(2), 455–474.

Cook, R. & Weisberg, S. (1991), 'Discussion of sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**, 328–332.

Cook, R. & Yin, X. (2001), 'Dimension reduction and visualization in discriminant analysis (invited, with discussion)', *Australia & New Zeland Journal of Statistics* **43**, 147–200.

Doocy, S. & Burnham, G. (2006), 'Assessment of socio-economic status in the context of food insecurity: Implications for field research', *World Health and Population* pp. 1–11.

Feeny, S., Mcdonald, L. & Posso, A. (2014), 'Are poor people less happy? findings from melanesia', *World Development* **64**, 448–459.

Forzani, L., García, R., Llop, P. & Tomassi, D. (2018), 'Supervised dimension reduction for ordinal predictors', *Computational Statistics and Data Analysis* **125**, 136–155.

García Arancibia, R., Depetris, E. & Rossini, G. (2015), 'From occasional consumption to alcohol abuse: Quantifying the socio-economic determinants in argentina', *Internatonal Journal of Development Research and Quantitative Techniques* **5**(1-2), 50–62.

Gertheiss, J. & Tutz, G. (2010), 'Sparse modelling of categorical explanatory variables', *The Annals of Applied Statistics* **4**, 2150–2180.

Hoque, S. (2014), 'Asset-based poverty analysis in rural bangladesh: A comparison of principal component analysis and fuzzy set theory', *SRI Papers, Sustainability Research Institute, University of Leeds* **59**.

Jackman, S. (2009), *Bayesian Analysis for the Social Sciences*, Wiley Series in Probability and Statistics, Wiley.
URL: *http://books.google.fr/books?id=QFqyrNL8yEkC*

Kamakura, W. & Mazzon, J. (2013), 'Socioeconomic status and consumption in an emerging economy', *International Journal of Research in Marketing* **30**, 4–18.

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

Li, B. & Wang, S. (2007), 'On directional regression for dimension reduction', *Journal of the American Statistical Association* **102**(479), 997–1008.

Li, B., Zha, H. & Chiaromonte, C. (2005), 'Contour regression: a general approach to dimensionreduction', *The Annals of Statistics* **33**(4), 1580–1616.

Li, K. (1991), 'Sliced inverse regression for dimension reduction (with discussion)', *Journal of the American Statistical Association* **86**, 316–342.

Li, K. C. (1992), 'On principal hessian directions for data visualization and dimension reduction: Another application of steinś lemma', *Journal of the American Statistical Association* **87**, 1025–1039.

Linting, M. & van der Kooij, A. (2009), 'Nonlinear principal components analysis with catpca: A tutorial.', *Journal of Personality Assessment* **94**(1), 12–25.

Mazzonna, F. (2014), 'The long-lasting effects of family background: A european cross-country comparison', *Economics of Education Review* **40**, 25–42.

Merola, G. & Baulch, B. (2014), Using sparse categorical principal components to estimate asset indices new methods with an application to rural south east asia, Conference Proceedings ABSRC, Rome, Italy.

Mokomane, Z. (2013), 'Social protection as a mechanism for family protection in sub-saharan africa', *International Journal of Social Welfare* **22**(3), 248–259.

Murasko, J. (2009), 'Socioeconomic status, height and obesity in children', *Economics and Human Biology* **7**, 376–386.

Richardson, D. & Bradshaw, J. (2012), *Family-Oriented Anti-Poverty Policies in Developed Countries*, Department of Economic and Social Affairs, Division for Social Policy and Development, United Nations, New York, New York.

Roy, K. & Chaudhuri, A. (2009), 'Influence of socioeconomic status, wealth and financial empowerment on gender differences in health and healthcare utilization in later life: Evidence from india', *Social Science and Medicine* **66**, 1951–1962.

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

Zhu, Y. & Zeng, P. (2006), 'Fourier methods for estimating the central subspace and the central mean subspace in regression', *Journal of the American Statistical Association* **101**(476), 1638–1651.