Using Sparse Categorical Principal Components to Estimate Asset Indices: new methods with an application to rural southeast Asia

GIOVANNI MARIA MEROLA* & BOB BAULCH**

* Xi'an Jioaton Liverpool University, ** International Food Policy Research Institute.

ABSTRACT: *Asset indices have been used since the late 1990s to measure wealth in developing countries. We extend the standard methodology for estimating asset indices using principal component analysis in two ways: by introducing constraints that force the indices to have increasing value as the number of assets owned increases, and by estimating sparse indices with a few key assets. This is achieved by combining categorical and sparse principal component analysis. We also apply this methodology to the estimation of per capita level asset indices. We show that the resulting asset indices improve the prediction and ranking of income both at household and per capita level using household survey data from northwest Vietnam and northeast Laos.*

Correspondence Address:Giovanni Merola, Xi'an Jioatong Liverpool University, 111 Ren'ai Road Suzhou, Suzhou Dushu Lake Science and Education Innovation District, Suzhou Industrial Park, Suzhou, P. R. China, 215123. Email giovanni.merola@xjtlu.edu.cn; Bob Baulch, International Food Policy Research Institute. 2033 K Street NW, Washington DC20006-1002, USA. Email b.baulch@cgiar.org.

1 Introduction

Since the late 1990s, researchers have used asset indices (AIs) as a relatively simple way to measure households' long term wealth or socioeconomic status in developing countries. The reason for using asset indices as a proxy for household income or expenditure stems from the well known difficulties associated with collecting comprehensive and reliable data on household income or expenditures (Deaton, 1997; Grosh and Glewwe, 2000), a desire in surveys focused on health or other issues to have a quick measure of household wealth (Gwatkin et al., 2007; Filmer and Pritchett, 2001) and the reduction of poverty and income dynamics due to measurement error (Carter and Barrett, 2006; Deaton, 1997; McKay and Perge, 2013). A recent review by Filmer and Scott (2012) analysed the results of a number of applications of asset indices and concluded that they are useful for the analysis of differences in health, education, fertility and child mortality.

In most applications asset indices are estimated by adapting methods designed for summarising continuous data to the categorical asset ownership and housing characteristic observed in household surveys. The most popular approach is to is to apply principal components (PCA) to dummy variables representing asset ownership, as originally proposed by Filmer and Pritchett (2001). Other methods used to compute AIs include factor analysis (Sahn and Stifel, 2000, 2003; Balen et al., 2010; Smits and Steendijk, 2015), polychoric PCA (Kolenikov and Angeles, 2004; Moser and Felton, 2007)

and multiple correspondence analysis (MCA) (Booysen et al., 2005; Smits and Steendijk, 2015).

Despite the widespread adoption of AIs, concerns remain about both the statistical validity of the way AIs are constructed and the interpretability of the results generated. One of the major drawbacks of AIs computed from dummy variables is that the intrinsic ordering of counts of assets cannot be retained. Therefore, the coefficients corresponding to owning a large number of an asset may be smaller than the coefficients corresponding to owning a smaller number of the same asset. This is both counter-intuitive and troubling for the use of asset indices as a measure of wealth. A similar argument can be made for housing characteristics which, when used for estimating wealth, can be made more informative by ordering their categories by their quality or cost.

Another drawback of AIs is that they lack parsimony: they are often defined by hundreds of coefficients, one for each number of the assets owned and each type of housing characteristics. Therefore, the contribution of an individual asset to the index cannot be determined. Understanding which assets and housing characteristics are the major drivers for the variation of wealth across households could be of great importance for studying its socioeconomic fabric, designing future surveys and cross-country comparisons.

This paper proposes improving on Filmer and Pritchett's approach

to computing AIs by including monotonicity constraints which force the coefficients of dummy variables to respect the ordering of their corresponding categories. This can be readily done by applying categorical PCA (CATPCA) (Gifi, 1990; Michailidis and de Leeuw, 1998) to household surveys data. CATPCA is analogous to multiple correspondence analysis with the addition of monotonicity constraints. In this paper we compute the CATPCA components by applying PCA to categorical variables scaled using aspect analysis (Mair and De Leeuw, 2010).

We also apply least squares sparse principal components analysis (SPCA, Merola, 2015) to the aspect scaled categorical variables to derive sparse principal components, which show the key drivers of variation across households using only a limited number of variables. This involves only a small loss of optimality while retaining the monotonicity constraints. Interpreting sparse AIs is much simpler than interpreting AIs defined as combinations of all the variables, because a few key variables that explain the most variance of the dataset can be quickly identified. As far as we are aware, this is the first time that CATPCA and SPCA have been used together to compute sparse components for categorical variables.

Finally, we use the scaled categorical variables to compute individual (per capita) level AIs from the asset counts for each household and aspect scaled housing categories divided by household sizes. We show that these AIs are superior to the standard ones both in predicting income and in

classifying income quintiles.

The paper is organised as follows: in the next section we give a brief methodological overview of the statistical techniques used for estimating AIs, including PCA, CATPCA and sparse PCA. In Section 3 we illustrate the estimation of AIs using CATPCA and sparse PCA using household survey data from northwest Vietnam and northeast Laos. Finally, in Section 4 we provide some concluding remarks and suggestions for future research.

2 Approaches to Estimating Asset Indices Using Household Survey Data

Household surveys record the number of assets owned by a household and, in most cases, the characteristics of the housing in which they live. Assets counts are discrete numerical data with an inherent numerical ordering but have nonlinear, skewed and heteroscedastic, distributions. Housing characteristics, are categorical variables with no inherent ordering or units of measure.

An AI is a linear combination of household survey observations which summarises the wealth of a household. If we let $x_j, j = 1, \ldots, p$ represent the asset ownership counts and housing characteristics recorded, the AI is defined as:

$$I = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p,$$

where the coefficients a_j are called *loadings*. The values of the AI (the *scores*) can be used as a proxy for the households wealth.

PCA is the oldest and most commonly used method to obtain one or more linear combinations of observed variables. The resulting linear combinations (called Principal Components, PCs) successively explain the maximum possible variance of the observed variables. The loadings are obtained as the eigenvectors of the covariance matrix of the observed variables. When the variables are scaled to unit variance PCA is carried out on the correlation matrix. Since continuous measures of association, like covariance and correlation, are either nonmeaningful or noncomputable for discrete or categorical data, different approaches for computing components of the household survey data using PCA data have been suggested.

In order to carry out PCA on household survey data Filmer and Pritchett (2001) converted the discrete variables on asset ownership into dummy indicators, each representing one of their categories or counts. Deriving the PCs from dummy variables has some drawbacks. First, a large number of dummy variables have to be introduced in the model. Second, coding each categorical variable with several dummy variables artificially inflates the total variance of the dataset. Therefore the percentage of variance explained by the components is severely underestimated (Abdi and Valentin, 2007). Third, the coefficients of the dummy variables are not constrained to reflect the order of counts and housing characteristics. This means

that coefficients corresponding to owning a large number of assets may be smaller than those corresponding to owning fewer assets. This lack of monotonicity of the coefficients was also noted by Moser and Felton (2007) and Wittenberg and Leibbrandt (2017) and is a serious problem because the AIs lose discriminating power and their coefficients are hard to interpret.

As an example of lack of monotonicity, Figure 1 shows the PC loadings of the dummy variables representing the ownership of bicycles and cellular (cell) phones computed on the Vietnamese and Laotian household survey data that will be used in Section 3. The loadings for cellular phones are nonmonotonic for both provinces as households owning more than one cellphone receive a lower AI score than those owning only one. In contrast, the loadings for bicycles are monotonic in Houaphanh but nonmonotonic in Thanh Hoa province.

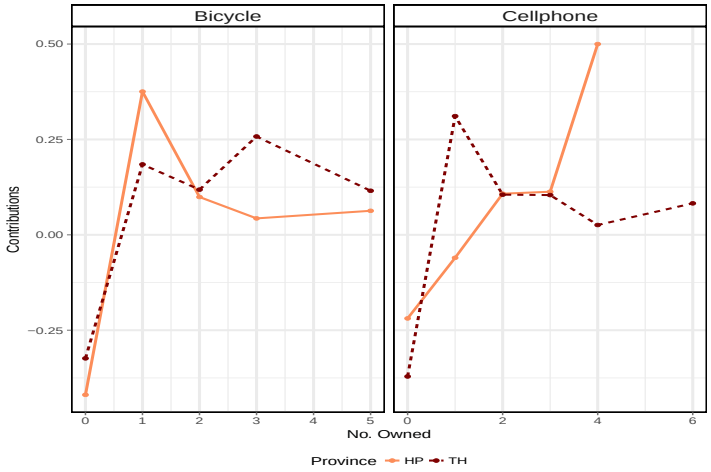

Figure 1: Loadings for the dummy variables representing the ownership of bicycles and cellular phones in Thanh Hoa (TH) and Houaphanh (HP).

Kolenikov and Angeles (2004) suggested a different method of computing PCA AIs using the matrix of polychoric correlations. Polychoric correlation measures the association between two ordinal variables without requiring the introduction of dummy variables and provides easy to interpret results. However, the use of polychoric correlation is somewhat controversial (see for example Agresti, 2002, p. 620, for a discussion). A practical concern regarding this approach is that the scores of the AI cannot be computed because the measures are nonnumerical and therefore cannot be multiplied by the loadings. This means that household wealth cannot be measured using this method.

Gifi (1990) shows that MCA is equivalent to carrying out PCA on the discrete variables transformed by assigning continuous numerical values to

each of their categories. Such *scaling*, known as homogeneity analysis (Gifi, 1990; Michailidis and de Leeuw, 1998), is determined by requiring that the first PC of the scaled variables explain the maximum possible variance of the dataset. In this sense, MCA is superior to Filmer and Pritchett's approach in which the correlation among the unscaled dummy variables is maximised.

MCA was extended using homogeneity analysis (Gifi, 1990; Michailidis and de Leeuw, 1998) and later aspect analysis (Mair and De Leeuw, 2010). One important advantage of these extensions over simple MCA is that the scalings can be restricted to maintain the monotonicity of the ordered categories, hence removing one of the major issue in computing PCA on ordered categories. Therefore, CATPCA improves on Filmer and Pritchett's approach by generating components that explain as much variance of the dataset as possible, while respecting the ordering of the categories Furthermore, the components are defined by only one loading for each variable which makes the results more interpretable.

As an illustration, Figure 2 provides a comparison of PCA loadings with monotonic CATPCA ones (converted to a dummy variables representation and rescaled) computed using the same assets and data as used in Figure 1. The first two panels show how the PCA loadings of owing a single cellphone in Houaphanh is larger than those of owing more than one cellphone. The same is true for bicycles in Thanh Hoa in the bottom right panel. In contrast, the CATPCA loadings are monotonically non-decreasing with the number

of cellphones and bicycles owned[1].

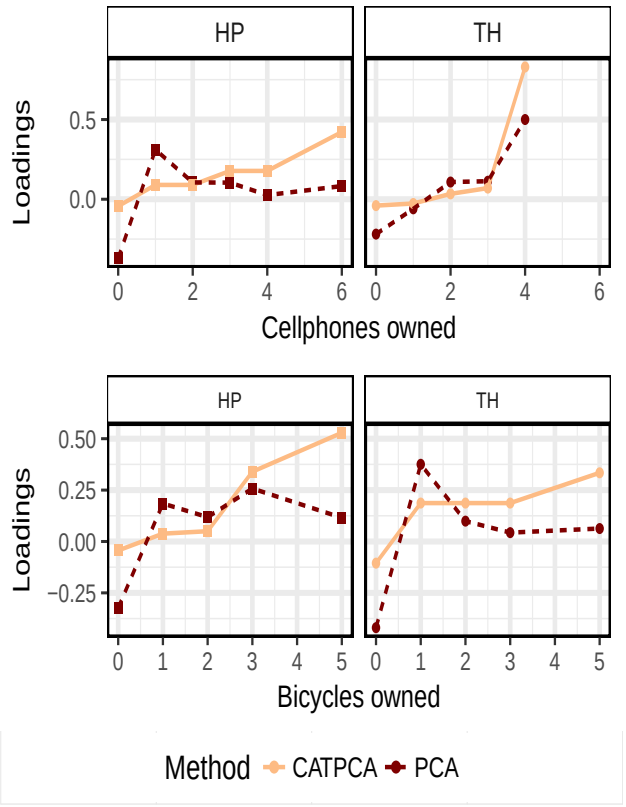

Figure 2: Comparison of CATPCA and PCA loadings for the number of bicycles and cellular phones owned for the Thanh Hoa (TH) and Houaphanh (HP) provinces.

2.1 Sparse Principal Component Analysis

A variant of PCA that has not yet been applied to the estimation of AIs is Sparse PCA (SPCA). SPCA aims to approximate the PCs of a data set with

[1]Plots comparing the different weighing for all items and housing characteristics may be found in Appendix A.3.

linear combinations of only a few of the variables. The number of nonzero loadings is referred to as the *cardinality* of the sparse component.

The motivation for SPCA is that interpreting standard principal components is not easy because they are combinations of *all* the observed variables. SPCA methods intend to replace the misleading practice (Cadima and Jolliffe, 1995) of *thresholding* the loadings by setting to zero small coefficients in an optimal way.

In recent years a number of methods for computing sparse PCs have been proposed (among others Zou et al., 2006; Moghaddam et al., 2006). Sparse components necessarily explain less variance than the corresponding PCs. As in all model selection problems, the tradeoff between parsimony and variance explained is complex and the computational cost of computing the solutions is high.

The sparse components computed by most SPCA methods are simply the PCs of a small subset of the observed variables (Moghaddam et al., 2006). As a result the sparse PCs explain well the highly correlated variables in the selected subset but ignore the variance of the variables that are not included (see Merola, 2015, for a discussion). Since an optimal SPCA solution cannot be found in reasonable time, these methods differ by how sub-optimal solutions to the problems are computed (see Trendafilov, 2013, for a review of these methods).

Merola (2015, 2018) developed least squares SPCA, a method in which

the sparse components explain the maximum variance of *all* the variables in the set. The optimisation problem is solved by a backward elimination algorithm or by projection. We will refer to the AIs computed by applying least squares SPCA to the variables scaled with aspect analysis as sparse CATPCA AIs.

2.2 Estimation of asset indices at the per capita level

AIs are typically computed at the household level. However, in many situations researchers are more interested in individual level measures, as in the case of poverty analysis. When the interest lays in per capita indicators it seems reasonable to consider per capita asset ownership. For example, a household with four members owning four cellphones should have a higher AI value than a household of size 10 owning four cellphones.

PCA on dummy variables cannot be computed for per capita data, so in most applications, the per capita level AI is simply derived by dividing the household level AI by the number of adult equivalent members. This method ignores the fact that the correlation structure of variables adjusted to per capita level is different from that of the original household level variables. Not surprisingly, Rutsein and Johnson (2004) found that this practice distorts the index distribution and its associations with health status and services, resulting in unreasonable results.

To compute the CATPCA AI at the per capita level, we adjust asset

counts and aspect analysis quantifications of the housing characteristics by dividing them by the respective household size. For the asset counts there is no need to compute aspect scalings because the asset counts divided by household sizes can be considered continuous monotonic measures. These AIs are then computed as the first PCs of the correlation matrices of these variables.

3 Estimation of Asset Indices in two rural provinces in Laos and Vietnam

To illustrate the advantages of sparse categorical PCA, we analyse data on assets and housing characteristics collected in two specialist household surveys conducted by Prosperity Initiative (PI) in two neighbouring provinces in northeast Lao PDR and northwest Vietnam: Houaphanh (HP) and Thanh Hoa (TH) in 2009 and 2010[2]. These five upland districts in which the household surveys were conducted were forested, mountainous and poor, and had similar agricultural and livelihood systems. However, Houaphanh is less densely populated than Thanh Hoa, while Thanh Hoa is generally more economically developed than Houaphanh. The surveys in

[2]Prosperity Initiative CIC was a UK-registered private interest company, which worked on promoting rural livelihoods and reducing poverty by enterprise development projects, particularly bamboo processing and handicrafts, in Cambodia, Lao PDR and Vietnam between 2006 and 2010.

both countries were representative, with housholds selected using a two-stage cluster sampling design and probability proportionate to size sampling: the Houaphanh survey surveyed 208 households in two districts while the Thanh Hoa survey covered 218 households in three districts. Both surveys collected data which allowed a comprehensive measure of household income including the value of consumption of own production to be estimated. Further details about sampling, survey and data collection can be found in the Appendix A.1.

The asset modules used in the surveys in both countries were virtually identical, and asked about households' ownership of different types of durable, productive, and other assets, together with how many of each asset were owned. This module is the sources of the data used to estimate the asset indices in this paper.

The assets included in the analysis for each province are listed in Table 1. They are classified into consumer durables, productive assets and means of transport. The table also shows the percentage of households owning each asset and the maximum numbers of each asset owned by a single household.

Table 1: Assets recorded in Thanh Hoa and Houaphanh by asset category.

		Than Hoa		Huaphanh	
Asset	Category	%	Max	%	Max
Cell	Durables	31%	6	51%	4
Cooking	Durables	48%	3	20%	4
DiningTable	Durables	3%	8	–	–
Elect	Durables	–	–	5%	3
Fan	Durables	79%	6	44%	11
Fridge	Durables	11%	2	19%	3
Generator	Durables	5%	2	28%	2
HHEquip1	Durables	27%	2	2%	2
HHEquip2	Durables	5%	3	–	–
Phone	Durables	53%	1	23%	2
Radio	Durables	9%	1	18%	2
SatDish	Durables	70%	2	68%	3
Sewing	Durables	–	–	30%	2
Sofa	Durables	17%	1	–	–
Stereo	Durables	7%	1	–	–
Stove	Durables	9%	2	–	–
TV	Durables	83%	2	74%	3
VCR	Durables	50%	1	54%	2
WtrHeater	Durables	–	–	5%	2
Buffalo	Productive	47%	7	46%	16
CartAnml	Productive	4%	1	3%	1
Cows	Productive	12%	9	38%	30
FeedGrind	Productive	7%	1	–	–
OtherProd	Productive	3%	22	6%	7
PestSpray	Productive	14%	1	–	–
Pump	Productive	24%	1	–	–
RiceMill	Productive	9%	1	58%	1
Thresh	Productive	17%	1	3%	1
Tractor	Productive	2%	2	55%	3
Bike	Transport	69%	3	85%	3
Car	Transport	2%	3	3%	2
PushBike	Transport	44%	5	36%	5
RowBoat	Transport	3%	1	–	–
Transp1	Transport	–	–	1%	2

'%' denotes the percentage of households owning the asset and 'Max' is the maximum number owned by a single household.

A comparison of the assets owned by households in Houaphanh and Thanh Hoa shows that while similar assets were owned in both provinces, the percentage of households owning different assets and the number owned varied significantly (Table 1). Except for a few assets (such as livestock, cellphones, motorcycles, sewing machines, and agricultural machinery),

Thanh Hoa generally had higher levels of asset ownership. However, because of livestock's role as a store of wealth in northern Eastern Lao, three times as many households owned cows in Houaphanh than in Thanh Hoa, and the maximum number of animals owned is also thrice as high there. Reflecting its greater remoteness, ownership of motorcycles and mobile phones were also more widespread in Houaphanh, while (fixed) telephones and bicycles were more common in Thanh Hoa.

To take advantage of the possibility of computing scalings respecting ordering, housing characteristics were ranked in increasing order of their approximate cost. The housing characteristics recorded and their categories ordered in increasing order of cost are shown in Table 2. This approach was also taken by Moser and Felton (2007).

Table 2: Housing characteristics recorded in Thanh Hoa and Houaphanh with their categories ordered by cost. The percentages observed is shown in parenthesis.

Huaphanh	
HouseType	temporary (6%), semi-prmnt (20%), 1 story (66%), multi-story/flat (8%)
Shared	yes (16%), no (84%)
Walls	bamboo (10%), wood (76%), concrete (5%), brick (9%)
Roof	grass (9%), wood (3%), metal sheets (31%), tiles (57%)
Floor	bamboo (13%), earth (8%), wood (55%), concrete (20%), tiles (4%), bamboo (13%)
Toilet	no toilet (10%), dry toilet (13%), flush toilet (76%)
WaterSource	Not piped (5%), public standpipe (78%), piped (17%)
Light	Combustion (5%), Generator (40%), Grid (55%)

Thanh Hoa	
HouseType	temporary (12%), semi-prmnt (71%), 1 story (14%), multi-story/flat (2%)
Walls	earth/other (6%), bamboo (19%), wood (45%), brick (30%)
Roof	straw/bamboo (39%), metal sheets (26%), tiles (28%), concrete (7%)
Floor	bamboo (45%), earth+lime/ash (12%), other (10%), cement + brick (18%)
Toilet	simple/no toilet (87%), latrine/suilabh (8%), flush toilet (4%)
WaterSource	container/other (4%), river/stream/pond (55%), well (38%), public standpipe (2%)
Light	noGrid (16%), grid (84%)

One-story houses made from wood are the most common in Houaphanh, while semi-permanent houses made from wood or brick are more prevalent in Thanh Hoa. More than half of households had tiled roofs in Houaphanh compared to just over a quarter in Thanh Hoa. In both provinces, bamboo or wood was the most common type of flooring. Over three-quarters of households in Houaphanh had flush toilets compared to just four per cent in Thanh Hoa. However, in Thanh Hoa, the vast majority of households had access to grid (mains) electricity, whereas almost half of households in Houaphanh had to use generators or lamps for their lighting.

3.1 CATPCA Asset Indices

The monotonic CATPCA AIs explain about 21 per cent of the variance of the data for Thanh Hoa and about 26 per cent for Houaphanh. The loadings of the CATPCA AIs, scaled to have the sum of their absolute values equal to one for both provinces, are shown together in Figure 3.1.1[3]. These scaled values can be interpreted as the percentage *contribution* of the corresponding variable to the AI. The few negative contributions present in both AIs have small values. In Thanh Hoa, the contributions of productive assets have lower values than for other asset classes; in Houaphanh, which is the poorer of the two provinces, assets in all classes have larger contributions.

One approach to interpreting the AIs' contributions is to threshold them,

[3]The numerical values of the loadings are shown in the online Appendix

that is ignore those which have absolute value lower than a given threshold and consider only the larger ones. However, it is difficult to identify a value that clearly cuts off "small" contributions from "large" ones. Furthermore, thresholding gives misleading results because if one or more variables are eliminated from the analysis, the loadings of the restricted AI would be different and should be recomputed. This is exactly what SPCA does, making the sparse solutions more appropriate for interpreting the loadings (see next section).

3.1.1 Sparse CATPCA asset index

We applied least squares SPCA to the correlation matrix of the variables scaled as before, requiring that the sparse AI explained at least 99% of the variance explained by the CATPCA AIs. This was achieved with 13 out of 34 assets for Houaphanh and 20 out of 38 assets for Thanh Hoa, as shown in Figure 3.1.1. In both provinces the sparse AIs are made up mainly of durable assets, with televisions, satellite dishes and fridges receiving the highest loadings in Houaphanh and fans, cooking stoves and fridges being the most important assets in Thanh Hoa. Housing characteristics (floor type and drinking water source in both provinces) also make substantive contributions to the sparse AIs.

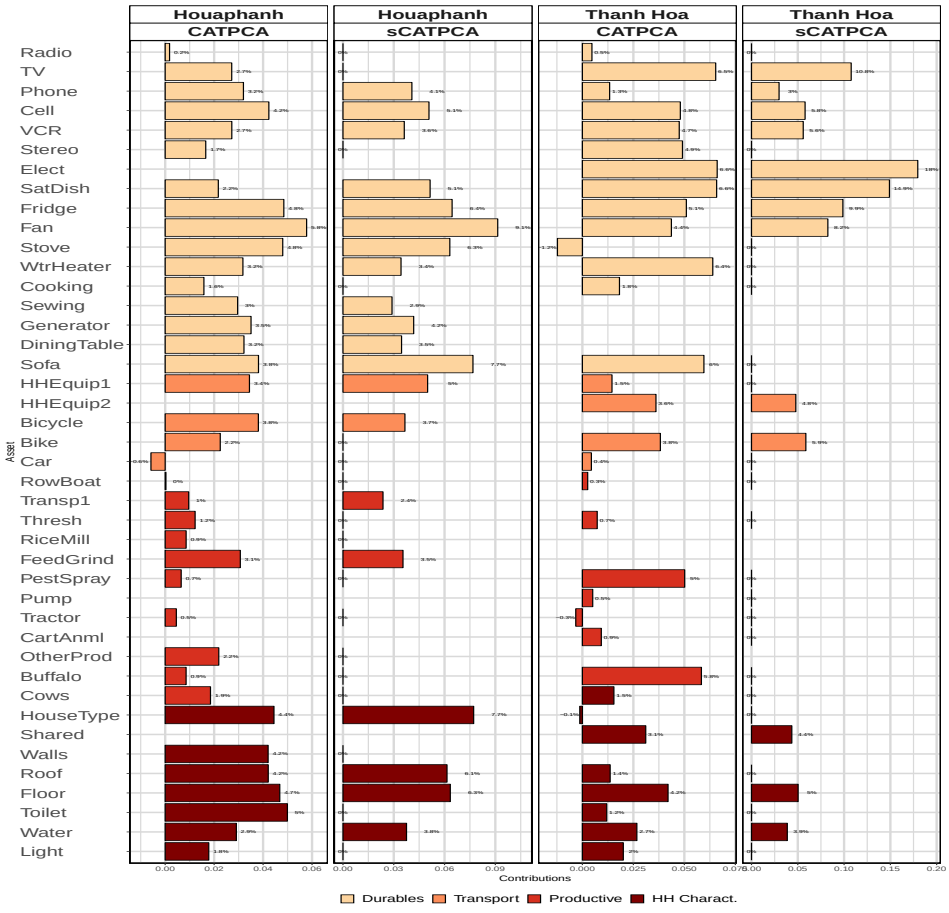

Figure 3: Full cardinality CATPCA and sparse sCATPCA AI at household level for Houaphanh (left) and Thanh Hoa (right) provinces.

The most evident change in the contributions due to sparseness is the reduction in the importance of productive assets. There are no productive assets in Houaphanh's sparse AI and only two (rice mills and pumps) contribute to Thanh Hoa's sparse index. This reflects the more diversified economic base in Thanh Hoa, when compared to poorer Houaphanh. The ownership of means of transportation are present in the sparse indices

for both countries; in more accessible and densely populated Thanh Hoa, ownership of both bicycles and motorbikes enter the sparse index. In Houaphanh, where there is lower population density and little public motorized transport, it is motorbikes and cars which enter into the sparse AI. It should be noted that the sparse AIs for both provinces eliminate the counterintuitive negative coefficients observed for some assets (such as generators, boats and carts) in the full CATPCA indices.

In Thanh Hoa province the sparse AI comprises more durables assets than in Houaphanh. In poorer Houaphanh province we find common assets, such as cellphones, satellite dishes and fridges; in richer Thanh Hoa we find more sophisticated assets, such as stoves, dining tables and other household equipment entering the index. As explained in Merola (2018), the SPCA approach we take is analogous to selecting a subset of the variables which explain, in a regression sense, the first PC with the required percentage of variability. This means that the variables excluded either have low correlation with the full cardinality AI or they are highly (multiply) correlated with other variables in the model and, therefore, are redundant for explaining the variance of the AI.

3.2 Per capita AIs

CATPCA AIs at the individual level have been computed as the first PC of the correlation matrix of asset ownership counts and aspect

scores for housing characteristics divided by household size. The per capita contributions, together with the corresponding household level contributions, are depicted in Figure 4[4].

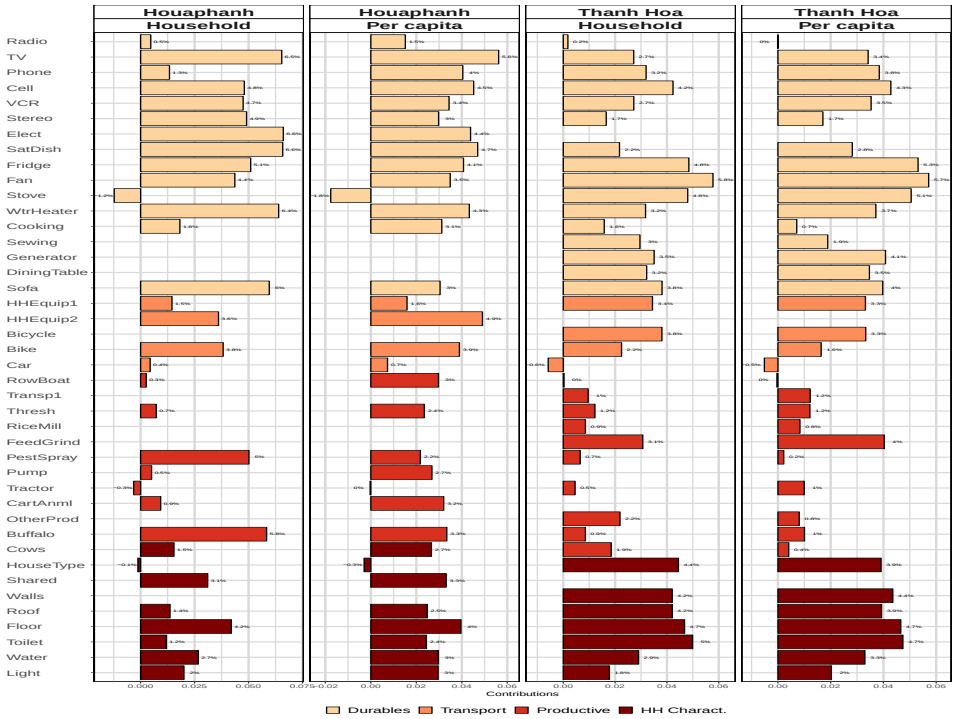

Figure 4: Per capita and household level CATPCA AI contributions for Houaphanh and Thanh Hoa province.

The contributions of the per capita AIs follow a similar pattern to the household level AIs in both provinces. This is surprising, as it would be expected that inexpensive assets used by individuals (such as bicycles or fans) would be less important when considered on a per capita basis. It should be also noticed that at the per capita level in Thanh Hoa owning

[4]The numerical values are available in the online appendix.

a water pump has a much larger contribution than other productive assets while at the household level owning cows and "other products" have large contributions. In Houaphanh, a larger number of productive assets have larger contributions at the per capita level than at the household level. To a lesser extent this is also true for housing characteristics.

3.2.1 Sparse per capita AIs

The contributions of the sparse CATPCA AIs computed at the individual level for both provinces are shown in Figure 5.

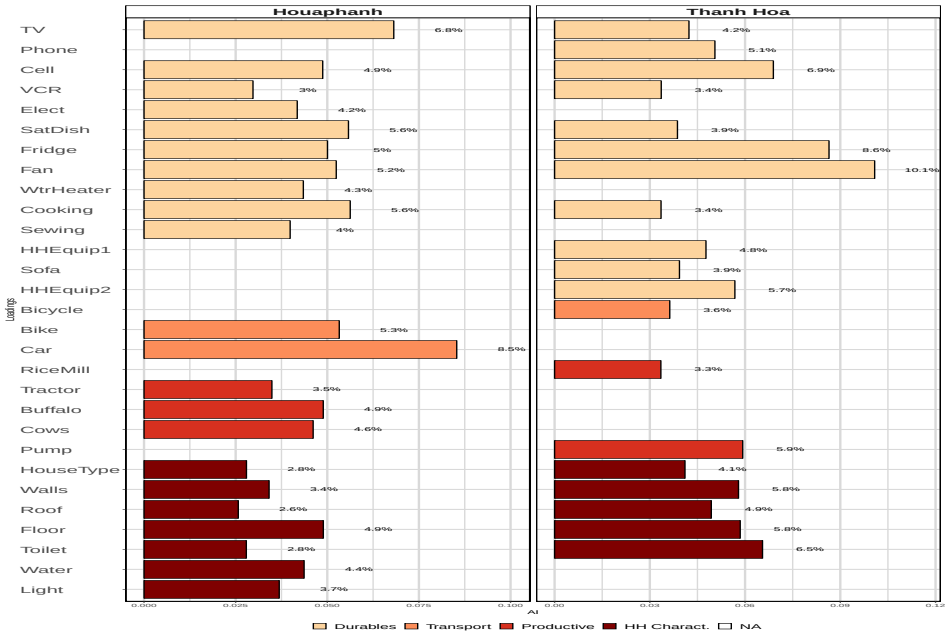

Figure 5: Contributions of the sparse per capita CATPCA AI.

The sparse AIs computed at the per capita level present similarities with

those computed at household level[5]. The Thanh Hoa and Houaphanh AIs present 19 and 22 nonzero contributions, respectively. One clear difference with the household level sparse loadings is that three productive assets in Houaphanh component now have non-zero loadings.

3.3 Assessing the asset indices

AIs are usually assessed by the extent to which they explain some welfare measures, most commonly expenditures (for example Filmer and Pritchett, 2001; Howe et al., 2009) or income (Filmer and Scott, 2012). Theoretically, AIs are neither proxies for expenditures nor for income. However, as household surveys in the education and health fields often do not record expenditures or income, AIs are sometimes used as proxies for welfare, though not always with satisfying results.

In evaluating different AIs, it should be remembered that PCA based methods find the components that best explain the variability of the observed data. These components will not necessarily be well suited for explaining welfare measures. So, if the assets surveyed are not good predictors of a measure, the estimated AI will not agree with it. Furthermore, if some assets are good predictors of a measure but have low correlation with the other variables in the data, the AI computed will likely not be a good predictor of the measure. Not surprisingly, Filmer and Scott

[5]The numerical values are shown in the online appendix.

(2012) therefore found that AIs were useful for measuring differences of certain welfare indicators but not of others. Wittenberg and Leibbrandt (2017) also argue that AI tend to exaggerate urban-rural differences by undervaluing rural assets (such as livestock), although this is unlikely to apply to our entirely rural sample. Recalling these conceptual difficulties, we now evaluate the different CATPCA AIs computed with respect to the measure of household income available in our data. The results are subsequently used as a benchmark for the performance of the CATPCA AIs and the PCA AI. The per capita PCA AI were computed by dividing the household level PCA AI scores by household sizes, as standard practice.

The box-plots in Figure 7, of the appendix A.3, show the distribution of the household level AIs scores (scaled to have zero minimum value and equal variance to achieve a meaningful comparison) for each quintile of household income in Houaphanh and Thanh Hoa, respectively. The distributions of the three indices are similar. However, the box-plots of the PCA AI are less separated than those of the CATPCA AIs, hence revealing less discriminating power.

All AIs show limited power to discriminate between households whose incomes are in the bottom two quintiles. This behaviour is usually explained by data clumping (McKenzie, 2005; Vyas and Kumaranayake, 2006; Howe et al., 2008). Such data clumping occurs when clusters of poor households own similar combinations of a few basic assets and have similar housing

characteristics, and so possess almost equal AI scores. Filmer and Scott (2012) explain the inability of standard AIs to discriminate between poorer households by the fact that they spend most of their income on food rather than on assets. This lack of discriminatory power does not depend on how the AIs are computed but on the range of assets included in the household survey (McKenzie, 2005).

In the following subsections, we first use regression analysis to evaluate the predictive precision of the different AIs. Then we consider how well the indices explain income, and how well they match actual income quintiles, at the household and per capita level.

3.3.1 Regression Analysis

As observed above, income and the AIs are right skewed, so the regression analysis gives better results without transforming income to its logarithm as commonly done in other applications.

In order to establish which assets are useful to explain income and to what degree they do, we eliminated nonsignificant coefficients through a backward stepwise variable selection regression (maximising Akaike's Information Criterion) of income onto the aspect scaled asset and housing variables at the household and per capita level. We use the result of these regressions as benchmarks for the regression of income on the different AIs. The results for Thanh Hoa and Houaphanh provinces are shown in Table 4

of the Appendix.

The assets selected by the stepwise regression seem intuitively more appropriate for measuring income than those assets having large contributions in the AIs. Expensive income generating assets, such as motor boats, cars, motorbikes, tractors and cows, are highly significant in the regression analysis but have small contributions in the AIs. Also, only a few housing characteristics are included in the solutions and then only at household level. However, the regression coefficients for a few assets and housing characteristics are difficult to explain because are negative. The adjusted R-squared values of the stepwise regressions are around 0.61, with the exception of the household level regression in Thanh Hoa – which has a value of 0.67. These quite high adjusted R-squared values show that the transformed asset and housing variables are useful for explaining income in the two provinces, both at the household and the per capita levels.

Table 3 compares the adjusted R-squared statistics for the stepwise regression of income with those from its regression on the CATPCA and PCA AIs, both at the household and the per capita level. As expected, the regression on the AIs yields lower adjusted R-squared statistics. The CATPCA AIs explain income markedly better at the per capita level than at the household level. The PCA AI explains income at the per capita level better in Thanh Hoa than in Houaphanh. The sparse CATPCA AIs explain slightly less variance of income than the corresponding full cardinality AIs.

Finally, the adjusted R-squared statistics for the regression of income on the PCA AIs are noticeably lower than corresponding ones obtained with the CATPCA AIs at both levels. This difference is larger at the per capita level, being 18% and 33% for Thanh Hoa and Houaphanh data, respectively.

Method	Than Hoa		Houaphanh	
	Household	Per capita	Household	Per capita
Step Reg	67.5%	61.4%	61.2%	61.2%
CATPCA	42.4%	50.9%	44.9%	52.4%
sparse CATPCA	41.2%	49.3%	44.2%	52.0%
PCA	36.9%	41.9%	42.3%	35.3%

Table 3: Adjusted R-squared statistics for the regression of income. The first row shows the statistic for stepwise backward regression on all variables, the remaining ones for the regression of income on the full and sparse and PCA AIs, respectively.

3.3.2 Classification matches and mistmatches

One of the main reasons why AIs are computed is to classify households or individuals into socio-economic groups. Figure 6 compares the performance the three different AIs considered at classifying units into quintiles. The top pair of diagrams show their performance in terms of regression adjusted R-squared; the second pair show the percentage of correct matches (same quintile of income and AI); and the third pair of diagrams show the percentage of gross mismatches (households or individuals classified two quintiles or more away by the AI from their respective income quintile[6]).

[6]For example units in the fifth quintile of income having AI score in the first or second quintiles of the AI.

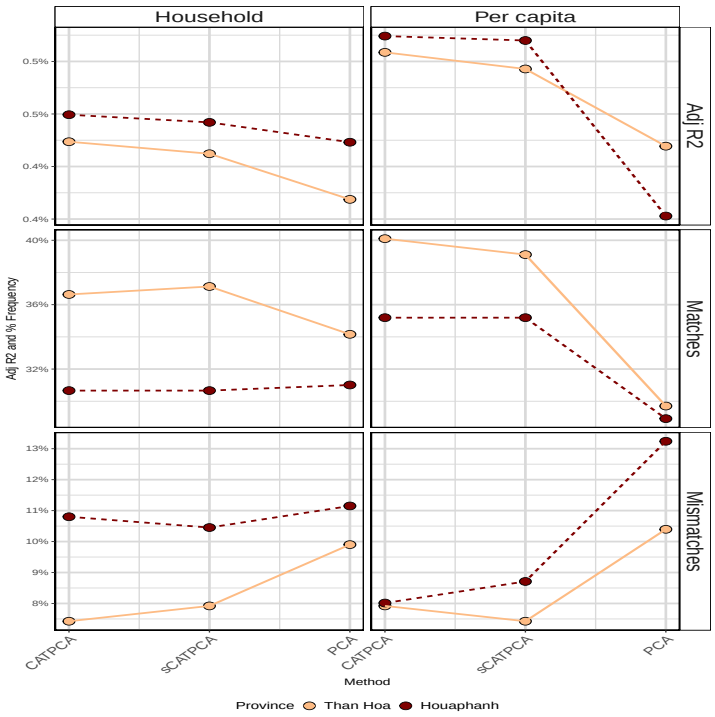

Figure 6: Adjusted R-squared statistics for the regression on income and percentages of matches and gross mismatches in quintiles of AI and Income

.

Clearly, PCA AIs perform worse than CATPCA AIs with respect to these measures, especially at the per capita level, where the loss in predictive power is more marked than at the household level. Again the full and sparse CATPCA AIs have similar performance. At the per capita level. stepwise regression gives worse classification than the CATPCA AIs, in some cases.

From the analysis presented above, we therefore conclude that using monotonic sparse and nonsparse CATPCA gives assets indices that are more interpretable and better at predicting income than standard PCA indices,

especially at the per capita level.

4 Conclusions

Asset indices have been used since the late 1990s as measures of long term socioeconomic status for households in developing countries. However, the methodologies currently used to compute asset indices do not guarantee that their scores increase with the number of assets owned and also lack parsimony. In this paper, we extend and improve the standard methodology for estimating asset indices in two ways: by introducing monotonicity constraints (which force the coefficients estimated to respect the ordering of the corresponding categories) and by using sparse principal components analysis (to reduce the number of variables on which the asset index is based). Both of these extensions are facilitated by the use of aspect analysis, which assigns optimal scalings to categorical variables representing asset ownership and housing characteristics. The adoption of aspect analysis also allows per capital level indices to be computed. To our knowledge this is the first time that these two methodologies have been used together

Asset indices built respecting the order of the ownership counts and the relative cost of housing characteristics are better able to measure long term wealth and socio-economic status. In addition to the intuitive appeal of using ordered loadings, the improvement from using CATPCA rather than standard PCA to estimate AIs has been demonstrated using household

survey data from neighbouring provinces in Laos and Vietnam. Not only do the AIs computed using CATPCA better explain income, but we also show that sparse indices give similar results to the full categorical ones. In our empirical application, sparse indices computed with between one-third and one-half of all assets surveyed, were able to explain 99 percent of the variance explained using all assets. The increase in parsimony and interpretability of AIs computed using sparse CATPCA makes them very attractive compared to standard AIs.

In the two provinces we studied in Laos and Vietnam, full and sparse CATPCA were also better at predicting household and per capita income than standard PCA. However, like standard AIs, both sparse and nonsparse CATPCA AIs show little power to discriminate between households whose incomes are in the bottom two quintiles. As explained in the text, this does not depend on how the AIs are computed but on the range of assets included in the household survey.

In conclusion, we believe that CATPCA and sparse CATPCA AIs represent an improvement over standard PCA methods for the estimation of asset indices in developing countries. At the time of writing, we are working on applying these indices to a wider range of household surveys, so as to be able to better identify the path forward for future research in this nascent field.

References

Abdi, H. and Valentin, D. (2007). Multiple correspondence analysis. In Salkind, N., editor, *Encyclopedia of measurement and statistics*. Sage Publications, Thousand Oaks (CA).

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 2nd edition.

Balen, J., McManus, D., Li, Y. S., Zhao, Z. Y., Yuan, L. P., Utzinger, J., Williams, G., Li, Y., Ren, M. Y., Liu, Z. C., Zhou, J., and Raso, G. (2010). Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of hunan province, china. *Ejacmerging Themes in Epidemiology*, 7(1):1–17.

Booysen, F., van der Berg, S., Burger, R., von Maltitz, M., and du Rand, G. (2005). Using an asset index to assess trends in poverty in seven sub-saharan countries. In *Proceedings of the International Poverty Centre Conference on Multidimensional Poverty, Brasilia 29-31 August, 2005*, volume `http://www.ipc-undp.org/conference/md-poverty/papers/Frikkie_.pdf`. published on-line.

Cadima, J. and Jolliffe, I. (1995). Loadings and correlations in the interpretation of principal components. *Journal of Applied Statistics*, 22(2):203–214.

Carter, M. and Barrett, C. (2006). The economics of poverty traps and persistent poverty: An asset-based approach. *The Journal of Development Studies*, 42(2):178–199.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

Filmer, D. and Pritchett, L. (2001). Estimating Wealth Effects Without Expenditure Data-or Tears: An Application to Educational Enrollments in States of India. *Demography*, 38(1):115–132.

Filmer, D. and Scott, K. (2012). Assessing asset indices. *Demography*, 49(1):359 – 392.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley, Chichester, England.

Grosh, M. and Glewwe, P. (2000). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study, Volume 2*. Washington, DC: World Bank.

Gwatkin, D. R., Rutstein, S., Johnson, K., Suliman, E., and Wagstaff, A. (2007). Socio-economic differences in health nutrition and population. bangladesh 1996/97 1999/2000 2004. Technical report, Washington DC

World Bank Human Development Network [2007]. available on-line at `http://bit.ly/1SiXIYb`.

Howe, L., Hargreaves, J., Gabrysch, S., and Huttly, S. R. A. (2009). Is the wealth index a proxy for consumption expenditure? A systematic review. *J Epidemiol Community Health*, 63(11):871 – 877.

Howe, L., Hargreaves, J., and Huttly, S. (2008). Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerging Themes in Epidemiology*, 5(1):1–14.

Kolenikov, S. and Angeles, G. (2004). The use of discrete data in pca: theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, 20:1–59.

Mair, P. and De Leeuw, J. (2010). A general framework for multivariate analysis with optimal scaling: The r package aspect. *Journal of Statistical Software*, 32(9):1–23.

McKay, A. and Perge, E. (2013). How strong is the evidence for the existence of poverty traps? a multicountry assessment. *The Journal of Development Studies*, 49(7):877–897.

McKenzie, D. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, 18(2):229–260.

Merola, G. (2015). Least squares sparse principal component analysis: a

backward elimination approach to attain large loadings. *Australia & New Zealand Journal of Statistics*, 57(3).

Merola, G. (2018). Projection sparse principal component analysis: an efficient least squares method. Submited for publication. Preprint available at `https://arxiv.org/abs/1612.00939`.

Michailidis, G. and de Leeuw, J. (1998). The gifi system of descriptive multivariate analysis. *Statistical Science*, 13:307–336.

Moghaddam, B., Weiss, Y., and Avidan, S. (2006). Spectral bounds for sparse pca: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pages 915–922. MIT Press.

Moser, C. and Felton, A. (2007). The construction of an asset index measuring asset accumulation in ecuador. Technical Report 87, Chronic poverty research centre working paper. available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1646417`.

Rutsein, S. O. and Johnson, K. (2004). The dhs wealth index. DHS Comparative Reports 6, ORC Macro, Calverton, Maryland, USA.

Sahn, D. and Stifel, D. (2000). Poverty comparisons over time and across countries in africa. *World Development*, 28(12):2123 – 2155.

Sahn, D. and Stifel, D. (2003). Exploring alternative measures of welfare in

the absence of expenditure data. *Review of Income and Wealth*, 49(4):463–489.

Smits, J. and Steendijk, R. (2015). The international wealth index (iwi). *Social Indicators Research*, 122(1):65–85.

Trendafilov, N. (2013). From simple structure to sparse components: a review. *Computational Statistics*, 28(4).

Vyas, S. and Kumaranayake, L. (2006). Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21(6):459–468.

Wittenberg, M. and Leibbrandt, M. (2017). Measuring inequality by asset indices: A general approach with application to south africa. *Review of Income and Wealth*, pages 706–730.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

A Additional Material

A.1 Survey details

To illustrate the advantages of sparse categorical PCA, we analyse data on assets and housing characteristics collected in two specialist household surveys conducted by Prosperity Initiative (PI) in two neighbouring

provinces in northeast Lao PDR and northwest Vietnam: Houaphanh (HP) and Thanh Hoa (TH)[7]. The districts in which PI was working in both provinces were forested, mountainous and poor, and are also significant producers of bamboo.

The surveys conducted by PI in both provinces contained three components: a farmer (i.e., household) questionnaire, a trader questionnaire, and a commune/village (i.e., community) questionnaire. The household questionnaire in both countries collected information on the location and demographic composition of households, the education and employment of household members, their ownership and use of agricultural land, crop, livestock and aquaculture production (including detailed information revenues and costs), bamboo harvesting, collection and sales, income from other sources, housing, and fixed assets and durable goods[8]. This information allowed a comprehensive measure of household income (including both monetary and in-kind income) to be estimated.

The survey conducted by Prosperity Initiative[9] contained three components: a farmer (that is, household) questionnaire, a trader questionnaire, and a commune/village (that is, community) questionnaire.

[7]Prosperity Initiative CIC was a UK-registered private interest company, which worked on promoting rural livelihoods and reducing poverty by enterprise development projects, particularly bamboo processing and handicrafts, in Cambodia, Lao PDR and Vietnam between 2006 and 2010.

[8]In addition, because of the importance of non-timber forest products (NTFP) to the income of some households there, the Houaphanh farmer questionnaire also contained an additional module on collection and sales NTFP.

[9]The team which designed and administered the questionnaire was led by the second author. The questionnaire was designed to conform with the best international practices for collecting household survey data (see Grosh and Glewwe, 2000).

Households were sampled according to a two-stage clustered design using probability proportionate to size (PPS sampling) in both countries, with villages in Houaphanh and communes in Vietnam being the primary sampling units. The Houaphanh survey covered 208 households in 27 villages in two districts (Sopbao and Viengxay), while the Thanh Hoa survey covered 218 households in 29 communes from three districts (Ba Thuoc, Lang Chanh and Quan Hoa). Although, administered in different languages (and scripts) by different survey teams, the questionnaires used to collect information from farmers were very similar in both countries. In particular, the asset modules used in both countries were virtually identical, and asked about households' ownership of different types of agricultural, durable, productive, and other assets, together the number owned and how many of each asset has been purchased in the last year. This module is the sources of the data used to estimate the asset indices in this paper.

For the estimation of the AIs, we eliminated three households in Thanh Hoa and six in Houaphanh which were recorded as owning few (three in Thanh Hoa and four in Houaphanh) different types of assets but had income above the median value. Assets surveyed that were owned by less than 2.5% households were grouped into other assets categories. As customary in household surveys, the questionnaire already contained generic '*other*' categories for asset types not included in the list. Since these categories often were recorded for fewer than 2.5% of the households, we merged them

with rarely owned assets. Assets were merged as follows: personal computer, water heater, sewing machine and washing machine were grouped into the household equipment 2 (*HHEquip2*) category. Car and motor boat were merged into a new asset type (*CarBoat*) because the two households that owned at least one car and the two that owned at least one motor boat were all in the top 10% income bracket. Tractor 2 Wheels and Tractor 4 Wheels were merged into *Tractor*. Other agricultural equipment, other animals, lathes and welding machines, sewing and weaving machines and other nonagricultural machinery were grouped into the new type *OtherProd*. Assets types that were not owned by any household were simply removed from the dataset.

A.2 Numerical Tables

Table 4: Coefficients of backward stepwise regressions of income on scaled asset and housing variables.

Thanh Hoa						Houaphanh					
Asset	AssetClass	Household level		Per Capita level		Asset	AssetClass	Household level		Per Capita level	
		Coeff	p-value	Coeff	p-value			Coeff	p-value	Coeff	p-value
Radio	Durables	2641.5	0.029	623.4	0.022	TV	Durables	-	-	273.5	0.025
TV	Durables	4503.5	0.000	-	-	Cell	Durables	906.4	0.092	238.2	0.030
Phone	Durables	1963.5	0.105	-	-	VCR	Durables	-	-	227.7	0.025
Cell	Durables	2013.1	0.171	464.1	0.162	Elect	Durables	1071.3	0.061	-	-
Stereo	Durables	-	-	-525.4	0.078	SatDish	Durables	5153.2	0.000	-	-
Stove	Durables	4862.7	0.001	1725.3	0.000	Fridge	Durables	-	-	236.1	0.032
Generator	Durables	-	-	-740.2	0.010	Fan	Durables	-1414.5	0.022	-249.0	0.024
DiningTable	Durables	-	-	502.8	0.079	WtrHeater	Durables	1157.9	0.056	-	-
Sofa	Durables	-	-	829.0	0.006	Cooking	Durables	-	-	193.6	0.073
HHEquip1	Durables	-	-	-715.2	0.032	Sewing	Durables	1517.8	0.001	-	-
HHEquip2	Durables	5008.1	0.000	1448.5	0.000	ExpHH	Durables	2521.7	0.004	-	-
Bicycle	Transport	-	-	503.0	0.099	Bike	Transport	2888.5	0.000	605.2	0.000
Bike	Transport	2652.4	0.066	809.7	0.008	Transp1	Transport	758.5	0.075	-	-
RowBoat	Transport	-	-	774.5	0.008	RiceMill	Productive	900.9	0.032	175.9	0.057
RiceMill	Productive	3004.5	0.049	-	-	Tractor	Productive	1169.6	0.006	176.7	0.054
FeedGrind	Productive	5214.4	0.001	1382.9	0.000	Buffalo	Productive	612.0	0.156	423.4	0.000
Pump	Productive	2903.7	0.033	1080.1	0.002	Cows	Productive	-2843.2	0.005	-	-
Tractor	Productive	6636.0	0.000	-	-	Floor	HH Charact.	-727.0	0.128	-	-
CartAnml	Productive	-2202.0	0.056	-	-						
Cows	Productive	6196.8	0.000	-	-						
HouseType	HH Charact.	3827.5	0.004	-	-						
Water	HH Charact.	-2009.2	0.162	-	-						
Light	HH Charact.	-3642.3	0.005	-	-						
Multiple R-squared		0.71		0.65				0.63		0.63	
Adjusted R-squared		0.67		0.61				0.61		0.61	

A.3 Figures

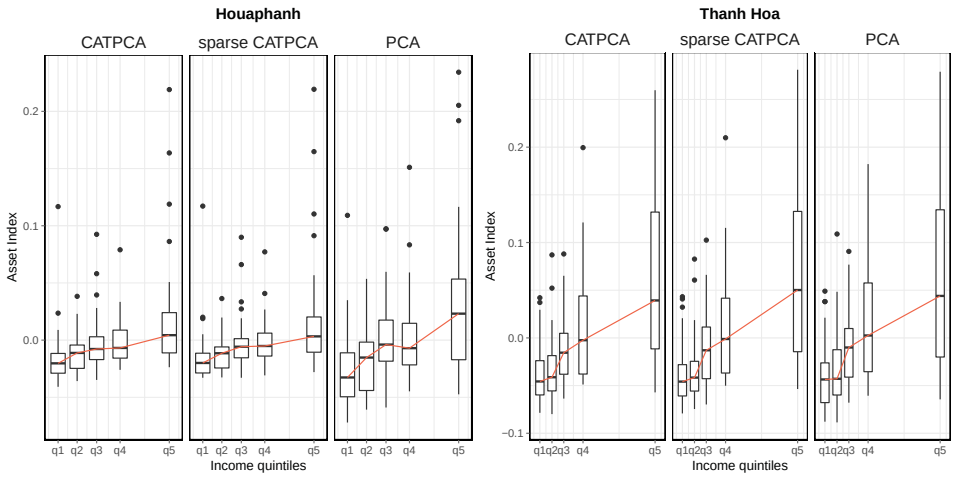

Figure 7: Full and sparse household level AIs plotted separately for different quintiles of income.

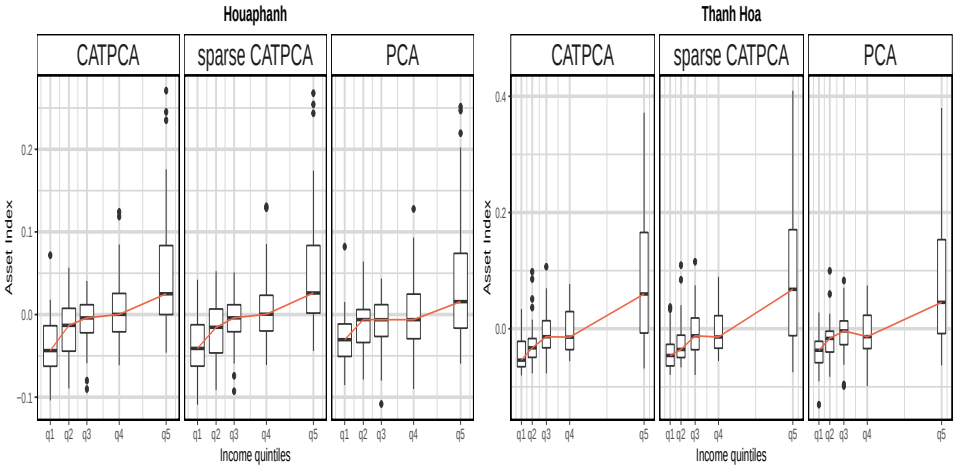

Figure 8: Full and sparse per capita level AIs plotted separately for different quintiles of income.