Hai-Anh Dang (World Bank) and Francois Bourguignon

Investigating Welfare Dynamics with Repeated Cross Sections: A Copula Approach

I. Introduction

Panel household surveys are indispensable for tracking household welfare dynamics over time. Yet, such surveys are rarely available, particularly for developing countries for various reasons. These can range from lack of financial resources (i.e. it is costly to implement panel surveys), and technical capacity (i.e., certain levels of technical expertise are required to maintain nationally representative panel surveys) to logistical challenges (i.e., in fragile and conflict contexts, it is difficult to implement surveys and/ or track households over time). Even where panel data are collected, such data do not often provide nationally representative data. For example, two middle-income countries, China and India, recently collected some panel data but these panel data are not commonly employed to provide poverty estimates. The surveys that are used in these countries for this purpose—the China Household Income Project (CHIP) survey and the National Sample Survey (NSS)—are both cross-sectional surveys (Dang and Carletto, 2018).

This data shortage has, in fact, been the main obstacle that hinders research on poverty mobility in developing countries.[1] More generally, researchers and policy makers face the same data challenge when trying to better understand the dynamics of other welfare outcomes other than poverty such as income mobility.

Recent statistical methods have been developed to overcome this data challenge, such that synthetic panels can be constructed using only two rounds of repeated cross sections (Dang et al., 2014; Dang and Lanjouw, 2013).[2] These synthetic panels have been validated against actual panel data and employed to study poverty transitions in a number of developing countries, including countries in Latin Americas (Ferreira *et al.,* 2013; Cruces *et al.,* 2015; Vakis *et al.,* 2015), Europe and Central Asia (Cancho *et al.,* 2015), the Middle East and North Africa (Dang and Ianchovichina, 2018), Sub-Saharan Africa (Dang and Dabalen, in press), and India (Dang and Lanjouw, 2018a and 2018b).[3] Most recently, Bourguignon, Moreno, and Dang (2019) further

[1] Still, the availability of cross section surveys should not be taken for granted, especially for poorer countries. A recent survey by Beegle *et al.* (2016) points out that just more than half (i.e., 27) of the 48 countries in Sub-Saharan Africa had two or more comparable household surveys for the period between 1990 and 2012. Even worse, Serajuddin et al. (2015) find that, over the period 2002- 2011, more than one-third (i.e., 57) of the 155 countries for which the World Bank monitors poverty data using the WDI database have only one poverty data point or no data at all.

[2] Bourguignon et al. (2004) provide an early attempt to construct synthetic panels to study poverty using more rounds of survey data.

[3] Researchers at international organizations including the UNDP, the Asian Development Bank, and the OECD have also applied these methods for analysis of welfare mobility (UNDP, 2016; Jha *et al.*, 2018; OECD, 2018); see also OECD (2015) for an application by the OECD to study labor transitions in richer countries.

extend this method in various directions to study income mobility and apply their method to data from Mexico.

In this paper, we build on the existing methods to construct synthetic panels in an alternative and more general way using copulas. Copulas require fewer parametric assumptions and have been widely used in other fields such as engineering or finance to provide a flexible estimate of the joint distributions of different marginal functions. [4] These copula-based synthetic panels allow us to offer more accurate estimates of general income (consumption) mobility, rather than just poverty and vulnerability mobility. Furthermore, we can offer estimates of various other absolute and relative mobility measures and indexes, such as income movement, positional movement, and non-anonymous growth incidence curves (GIC). In terms of modelling techniques, we offer both semi-parametric and non-parametric approaches (i.e., theoretical and empirical copulas). Our method are also straightforward to implement.

We will validate our proposed method using both actual panel data and repeated cross sections from several countries at different income levels and from different geographical regions, such as Vietnam and the US. We will analyze several latest survey rounds of the Vietnam Household Living Standards Survey (VHLSS) and the Panel Study of Income Dynamics (PSID) that span the early to mid-2010s. We will also supplement our analysis with panel data from other countries including Bosnia-Herzegovina, Indonesia, Lao PDR, and Peru.

II. Preliminary Estimation Results

We start first with showing estimates for poverty transitions against those based on the actual panels—or the "true" estimates—in Table 1 for the two years 2016 and 2018 of the VHLSS. For a start, we employ the Gaussian copula for analysis. The synthetic panel estimates encouragingly fall within the 95 percent confidence interval (CI) of the true estimates in all four cases. For example, the chronic poverty rate using the synthetic panel is 9.7 percent, which is not statistically significantly different from the true chronic poverty rate of 9.9 percent with a CI of 0.8 percent (Table 1, first row). In fact, the synthetic panel estimates even lie within one standard error of the true estimates for the majority of the cases (i.e., three out of four). Another validation indicator, the mean coverage rate that measures the common overlap between the CIs of the synthetic panel estimates and those of the actual panel (Dang and Lanjouw, 2013) is also reasonably strong. It is estimated to be 85 percent. Furthermore, for two of the four estimates, the CIs around the synthetic panel estimates fall completely inside those of the actual estimates.

We go beyond the two by two poverty transition in Table 1, and further examine in Table 2 the more general five by five transition matrix for the consumption quintiles. Four-fifths (i.e., 20 out of 25) of the synthetic panel estimates for the inner cells are not statistically significantly different from those based on the actual panels. These estimates are marked in bold letters.

[4] See, e.g., Nelsen (2006) and Trivedi and Zimmer (2007) for an introduction to copulas.

Furthermore, more than half of the former fall within one standard error of the latter, which are marked with a star.

Table 3 produces estimates for another index of consumption mobility, which is the median consumption growth rate over time for households in the four groups earlier defined in Table 1. The growth rates of consumption are more difficult to estimate than the poverty transitions in Table 1, because the former require more accuracy with households' exact levels of consumption rather than just scoring whether households fall below a given poverty line as with the latter. Yet, Table 3 provides encouraging estimation results with three of four estimates and two of four estimates respectively falling within the 95 percent CI and one standard error of the true estimates.

Again, we generalize the estimation results in Table 3 and plot the non-anonymous growth incidence curve for Vietnam over the period 2006-08 in Figure 1. Our estimates perform reasonably well, with the synthetic panel estimates (solid green line) mostly falling inside the 95 percent CI (gray area) around the true estimates (dotted red line).

III. Further Analysis

We plan to further refine the proposed copula methods and estimation results above along different directions. These include the following

i. examine robustness of estimation results using different copulas as well as the non-parametric empirical copula, including testing for theoretical and empirical differences with different copulas

ii. conduct heterogeneity analysis for different population groups, including by gender, education levels, ethnicity, urban/rural residence

iii. estimate other measures of absolute and relative mobility, including a five by five transition matrix where the thresholds are fixed using only the first-year or second-year thresholds (rather than in both years with Table 2), and other measures such as Fields-Ok index and mobility as equalizer of longer-term inequality.

iv. investigate whether, and to what degree, that measurement errors may impact our estimation results for mobility

Table 1: (Unconditional) Poverty Transitions Based on Synthetic Data for Two Periods, Using Gaussian Copula (Percentage)

Poverty Status	**Vietnam**
First Period & Second Period	**2006-08**

	Actual Panel	**Synthetic Panel**
Poor, Poor	9.9	9.7
	(0.8)	(0.6)
Poor, Nonpoor	5.9	4.9
	(0.5)	(0.4)
Nonpoor, Poor	4.9	5.1
	(0.5)	(0.4)
Nonpoor, Nonpoor	79.3	80.3
	(1.0)	(0.8)
Goodness-of-fit Tests		
Within 95% CI	4/4	
Within 1 standard error	3/4	
Mean coverage (percent)	85.2	
Coverage of 100%	2/4	
N	2723	3701

Table 2: Consumption Dynamics for Two Periods, Using Gaussian Copula, Vietnam 2006-2008 (Percentage)

			2008					
			Poorest	**Quintile 2**	**Quintile 3**	**Quintile 4**	**Richest**	**Total**
Panel A: True Panels	**2006**	**Poorest**	12.7	4.7	1.7	0.6	0.2	19.7
			(0.8)	(0.4)	(0.3)	(0.2)	(0.1)	(0.9)
		Quintile 2	4.8	7.5	4.6	2.0	0.6	19.6
			(0.4)	(0.6)	(0.5)	(0.3)	(0.1)	(0.9)
		Quintile 3	1.8	5.2	6.9	4.6	1.5	20.0
			(0.3)	(0.5)	(0.5)	(0.5)	(0.2)	(0.9)
		Quintile 4	0.6	2.0	5.0	7.8	4.8	20.2
			(0.2)	(0.3)	(0.5)	(0.6)	(0.5)	(0.9)
		Richest	0.1	0.6	1.8	4.9	12.9	20.5
			(0.1)	(0.2)	(0.3)	(0.5)	(0.7)	(0.8)
		Total	20.0	20.0	20.0	20.0	20.0	100
			(1.0)	(0.9)	(0.9)	(0.9)	(0.9)	
			2008					
			Poorest	**Quintile 2**	**Quintile 3**	**Quintile 4**	**Richest**	**Total**
Panel B: Synthetic Panels	**2006**	**Poorest**	**13.7**	**4.7***	**1.4**	0.2	**0.0**	20.0
			(0.3)	(0.3)	(0.2)	(0.1)	(0.0)	(0.0)
		Quintile 2	**4.7***	**8.0***	**5.3**	**1.8***	0.2	20.0
			(0.3)	(0.3)	(0.3)	(0.2)	(0.1)	(0.0)
		Quintile 3	**1.4**	**5.3***	**7.1***	**5.1**	**1.1**	20.0
			(0.2)	(0.3)	(0.3)	(0.3)	(0.2)	(0.0)
		Quintile 4	0.2	**1.9***	**5.1***	**8.2***	**4.6***	20.0
			(0.1)	(0.2)	(0.3)	(0.4)	(0.3)	(0.0)
		Richest	**0.0**	0.2	1.1	**4.7***	**14.1**	20.0
			(0.0)	(0.1)	(0.2)	(0.3)	(0.3)	(0.0)
		Total	20.0	20.0	20.0	20.0	20.0	100
			(0.0)	(0.0)	(0.1)	(0.1)	(0.0)	

Note: Synthetic panels are constructed from cross sections for Vietnam. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design. All numbers are weighted using population weights. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. Joint probabilities are shown. Estimates based on the synthetic panels that fall within the 95% CI and one standard error of those based on the actual panels are shown respectively in bold and in bold with a star "*".

Table 3: Median Consumption Growth for Two Periods, Using Gaussian Copula (Percentage)

Poverty Status	**Vietnam**	
First Period & Second Period	**2006-08**	
	Actual Panel	**Synthetic Panel**
Poor, Poor	3.6	3.5
	(0.3)	(0.3)
Poor, Nonpoor	9.1	8.5
	(0.4)	(0.2)
Nonpoor, Poor	-1.5	-1.5
	(0.3)	(0.3)
Nonpoor, Nonpoor	3.1	2.9
	(0.1)	(0.1)
Goodness-of-fit Tests		
Within 95% CI	3/4	
Within 1 standard error	2/4	
Mean coverage (percent)	75.9	
Coverage of 100%	1/4	
N	2723	3701

Figure 1: Non-anonymous Growth Incidence Curve, Vietnam 2006-2008