Latent Heterogeneity in the Marginal Propensity to Consume*

Daniel Lewis
FRBNY

Davide Melcangi
FRBNY

Laura Pilossoph†
FRBNY

November 8, 2019

Abstract

We estimate the distribution of marginal propensities to consume (MPCs) using a new approach based on the fuzzy C-means algorithm (Dunn (1973), Bezdek (1981)). The algorithm generalizes the K-means methodology of Bonhomme and Manresa (2015) to allow for uncertain group assignment and to recover unobserved heterogeneous effects in cross-sectional and short panel data. We extend the fuzzy C-means approach from the cluster means case to a fully general regression setting and derive asymptotic properties of the corresponding estimators by showing that the problem admits a GMM formulation. We apply the estimator to the 2008 tax rebate and household consumption data, exploiting the randomized timing of disbursements. We find a considerable degree of heterogeneity in MPCs, which varies by consumption good, and provide evidence on their observable determinants, without requiring ex ante assumptions on such relationships. Our aggregated heterogeneous results suggest the partial equilibrium consumption response to the stimulus was twice as large as what is implied by homogeneous estimates.

Keywords: Marginal Propensity to Consume, Consumption, Tax Rebate, Heterogeneous Treatment Effects, Machine Learning, Clustering, C-means, K-means
JEL Codes: D12, D91, E21, E32, E62

*This paper was previously circulated as "A New Approach to Estimating Heterogeneous Effects and the Distribution of the Marginal Propensity to Consume". The views expressed herein are solely those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York or the Federal Reserve System. We thank Rene' Chalom and Meghana Gaur for excellent research assistance. For helpful comments, we also thank Sara Casella, Richard Crump, Marco Del Negro, Keshav Dogra, Domenico Giannone, Simon Gilchrist, Isaac Sorkin, Mary Wootters, as well as various seminar and conference participants.
†daniel.lewis@ny.frb.org, davide.melcangi@ny.frb.org, laura.pilossoph@ny.frb.org.

1 Introduction

Recent work highlights the importance of heterogeneity in marginal propensities to consume (MPCs) out of transitory income shocks for fiscal policy, the transmission of monetary policy, and welfare.[1] Nonetheless, despite their importance, estimates of the distribution of MPCs are fairly elusive. Even with plausibly identified transitory income shocks, estimating individual-level MPCs requires panel data with long horizons, which are typically not available; it also usually requires the unappealing assumption that an individual's MPC is time invariant.[2] The existing literature, therefore, has followed one of two avenues: estimating a fully structural model and simulating a distribution of MPCs, or grouping observations by some presupposed observable characteristics and estimating group-specific MPCs out of transitory income shocks.[3,4] However, because both of these approaches require taking a stance on the source of MPC heterogeneity, they may either fail to uncover the true degree of heterogeneity, miss other relevant dimensions of heterogeneity that predict an individual's MPC, or both.

In this paper, we propose a new way to estimate the distribution of MPCs directly. We introduce a fuzzy C-means-based estimator (Dunn (1973), Bezdek (1981)) which jointly (i) groups households together that have similar latent consumption responses to the 2008 tax rebate and (ii) provides estimates of the MPCs within these groups. More specifically, the algorithm takes a standard regression of consumption changes on controls and the tax rebate receipt (Johnson et al. (2006), Parker et al. (2013)), but allows the coefficient on the rebate to be heterogeneous across groups; the groups as well as their rebate coefficients are jointly estimated so as to minimize a particular objective based on regression residuals. The approach is appealing because it allows us to estimate the unconditional distribution of MPCs directly first, without taking a stand on correlates of the distribution. Moreover, it does not require the assumption that an individual's MPC is time-invariant, or, in fact, any panel structure. We can therefore "let the data speak" by investigating *ex post* which observables predict the uncovered individual MPCs, including time varying household characteristics.[5] Indeed, we find a considerable degree of heterogeneity, and document a

[1] The MPC distribution is a crucial object in Heterogeneous Agent New Keynesian (HANK) models of monetary policy (see Kaplan et al. (2018)). For example, Auclert (2019) shows that the response of aggregate consumption to monetary policy shocks depends on the covariance of the distribution of MPCs with the cyclicality of income, net nominal position, and unhedged interest rate exposure.

[2] Nearly all theories of MPC heterogeneity have some form of state dependence.

[3] For the former, see for instance, Kaplan and Violante (2014) and Carroll et al. (2017).

[4] Fagereng et al. (2016) exploit lottery randomized winnings to identify transitory income shocks, and subsequently group observations on observables to estimate group-level MPCs. See also Johnson et al. (2006), Kaplan et al. (2014), Parker et al. (2013), and Crawley and Kuchler (2018).

[5] Other papers have used the "reported preference" approach, drawing MPC heterogeneity directly from

robust and significant positive relationship between the MPC and the average propensity to consume (APC), total income, and the presence of a mortgage on a household's balance sheet.

The approach we develop builds heavily on the K-means algorithm framework recently studied by Bonhomme and Manresa (2015), but in a cross-sectional or short-panel setting rather than a long panel data setting. Heuristically, the K-means algorithm begins by randomly assigning households to groups, then estimates heterogeneous regression coefficients, and finally reassigns households to groups by minimizing a residual-based objective until convergence. We consider instead the more general fuzzy C-means approach, which allows uncertain group assignment via continuous weights, rather than a binary assignment. Capturing this uncertainty means the algorithm is better suited to our cross-sectional environment, and short-panel data more broadly, which we corroborate in various simulation exercises. "Hard" K-means (HKM) remains a limiting case of fuzzy C-means (FCM).

First, to motivate the use of FCM, we show analytically that it can have smaller bias than HKM when $T = 1$, even in a simple cluster means case. We further show that there always exists some parameterization such that FCM is unbiased in this setting. Then, we extend the results of Yang and Yu (1992) and Yang (1994), who study the asymptotic properties of FCM for cluster means, to a fully general regression model. We start by showing that the FCM regression problem with simultaneously-determined weights is equivalent to a single-step nonlinear objective function. As a key contribution, we argue that this objective function fits naturally into the GMM framework (Hansen (1982)). Computationally, this eliminates the need for an iterative re-weighting algorithm, and, theoretically, it allows us to characterize the asymptotic distribution of the resulting estimator. We offer a further extension to two-stage least squares (TSLS) to accommodate the use of instrumental variables, providing a novel way that machine learning techniques can be exploited in the second-stage of IV estimation.[6] Based on our results, the fuzzy C-means approach is well-suited for a wide variety of economic settings with cross-sectional or short-panel data.[7]

Our estimator is attractive for various reasons. First, it is more flexible than standard clustering approaches, because each individual is assigned a non-binary weight on the

responses to survey questions. Recent examples include Jappelli and Pistaferri (2014) and Fuster et al. (2018).

[6]This differs from how machine learning techniques have generally been employed in IV settings, as a prediction tool to develop a strong instrument in the first stage (e.g., Belloni et al. (2012)).

[7]Further extensions to time-series and nonlinear regression context, while outside the scope of this paper, appear straightforward.

estimated group-specific MPCs. This explicitly accommodates the uncertain assignment inherent to common cross-sectional or short-panel data. The researcher can calibrate this "fuzziness" via a tuning parameter, with "hard" K-means as a limiting case. Second, it is computationally fast, because it can be solved nonlinearly without resorting to iterative procedures, which can be prohibitively costly in realistically large datasets often encountered in applied settings. Third, the estimators have standard asymptotic distributions, with analytical standard errors that perform well in simulations, when bootstrapping would be computationally burdensome. We demonstrate these properties in various simulation exercises which vary the "fuzziness" parameter, the number of groups, and the degree of cluster separation. We show that our FCM approach has advantages over HKM when the data are not well separated, despite the fact that HKM more accurately represents the data generating process. These results are consistent with the analytical comparison we derive for the simple cluster means case.

We apply the FCM estimator to study heterogeneity in the MPC using the 2008 Economic Stimulus Act, and we uncover a considerable degree of heterogeneity. Households span the whole spectrum of propensities, from nearly no response to the receipt of the rebate, to propensities of 1 and even slightly above, depending on the consumption good studied. This suggests that some households are severely constrained, consuming the rebate in its entirety. The vast majority of individuals, despite arguably not being currently constrained, display a positive MPC. Our results are consistent across different specifications and sample restrictions. Instrumenting the rebate with an indicator for its receipt, as in Parker et al. (2013), leaves the results qualitatively unchanged and in fact increases the estimated heterogeneity in MPCs. The same is true when we exclude from the sample households that never received a rebate, or when we include lagged values of the rebate to control for persistent effects of the rebate receipt.

We then show how the MPC distribution varies across consumption goods. A large share of households do not consume additional nondurable goods out of the transitory income shock, consistent with what would be expected from the perspective of the permanent income hypothesis (PIH, Friedman (1957)). Moreover, households at the right of the distribution consume a smaller fraction of the rebate in nondurables than in total expenditures. Furthermore, more than 75% of households do not adjust their durable consumption in response to the tax rebate. The households that do purchase new durables, however, display an MPC close to 1. These findings are consistent with the discreteness and infrequency of durable purchases, resulting in lumpy adjustments. Correlating the household level MPCs across these consumption categories, we find positive, yet small,correlation between MPCs out of durable and non-durable goods.

Having characterized the distribution of marginal propensities to consume, we describe its main drivers. We document that many observable characteristics are individually correlated with household MPCs, but only three of them, however, are robust to the inclusion of other controls. First, high-income households have greater propensities to consume. This result crucially hinges on total income. That is, the result holds including financial and business income, but does not hold for salary earnings. Second, having a mortgage is associated with displaying a higher MPC out of total expenditures. Third, a household's MPC and APC are positively correlated. We regard this result as particularly useful for disciplining macro models of household consumption and savings, as it is easy to compute expenditure rates, both in structural models as well as in the data. Finally, our best array of observable predictors is able to explain only 13% of the variance in estimated MPCs. This suggests that a relevant portion of MPC heterogeneity might be driven by latent, unobserved household traits. Such heterogeneity could never be recovered splitting the sample by observable characteristics and estimating within-subsample homogeneous MPCs, as typically done in the literature.[8]

Finally, correctly accounting for MPC heterogeneity also matters for the aggregated consumption effects of the fiscal stimulus. We show that the sample average of our estimated heterogeneous responses is larger than the homogeneous marginal propensity to consume. When considering the cumulated heterogeneous responses over two quarters, the aggregated response from our distribution is almost twice as large as its homogeneous counterpart. While still a partial-equilibrium object in nature, this result suggests that correctly accounting for heterogeneity is important in order to correctly evaluate the impact of the 2008 fiscal stimulus.

This paper relates to other approaches proposed to estimate heterogeneous responses, besides Bonhomme and Manresa (2015) as described previously. Bonhomme et al. (2017) consider a two-step grouped-fixed effects estimator, which classifies observations into groups via HKM in the first stage, and estimates group-specific heterogeneity in the second stage. First, our algorithm performs a joint estimation as in Bonhomme and Manresa (2015) rather than a two-step procedure, as the cross-sectional structure of our data does not make it well-suited for a sequential strategy. We extensively discuss the relationship between our FCM approach and HKM in the following sections. The clustering approach is also distinct from other machine learning methods used to recover heterogeneous ef-

[8]Preference heterogeneity might be elicited in survey questions. Using Nielsen panel data, Parker (2017) finds that the MPC out of the tax rebate is indeed strongly correlated with a self-reported measure of impatience. Aguiar et al. (2019) use a two-asset model and the panel dimension of the PSID to show that heterogeneity in discount factors and inter-temporal elasticities of substitution play a major role in explaining MPC heterogeneity.

fects (e.g., random forests, neural networks) as in Chernozhukov et al. (2017), since they rely on using a wide array of observables to characterize heterogeneity. In our context, which is typical of many datasets, few observables are available; moreover, we seek to recover latent heterogeneity that may be entirely unrelated to observables.

Our approach is also distinct from that of Misra and Surico (2014), who study heterogeneous responses to the rebate using quantile regressions. Quantile regressions face limitations in the present setting. If the true heterogeneity is continuous, then the way in which it is discretized - and thus the results - is potentially driven by the researcher's choice of quantiles. If instead the heterogeneity is discrete (for example, a grouped structure), the researcher must rely on guessing the correct quantiles to line up with the distribution of groups. Moreover, if there are multiple dimensions of heterogeneity (whether continuous or discrete), the researcher must specify enough quantiles to accurately characterize their joint distribution. Again, little guidance is available for the choice of quantiles, and if too few are specified, the distribution recovered is likely to be biased, particularly across dimensions of heterogeneity. In contrast, clustering algorithms are equipped with theoretically-motivated model-selection tools to guide the researcher to choose the correct number of discrete groups or appropriately discretize continuous heterogeneity. Moreover, the size of each group is left unrestricted.

The rest of the paper proceeds as follows. In Section 2, we formulate the problem at hand and derive the FCM grouped marginal effects estimator. We extend the FCM from the cluster means case to a fully-general regression setting, as well as instrumental variables regression, and derive asymptotic properties of the corresponding estimators by showing that the problem admits a GMM formulation. The simulation studies presented in Section 3.2 demonstrate the performance of our FCM algorithm. We describe our empirical strategy based on the 2008 tax rebate in Section 4, and provide estimates of the distribution of MPCs for various consumption categories in Section 5. Section 5.3 discusses observable characteristics which correlate with the estimated MPCs. Section 6 aggregates the estimated household MPCs to arrive at a partial equilibrium effect on aggregate consumption and Section 7 concludes.

2 The Fuzzy C-means grouped marginal effects estimator

The fuzzy C-means algorithm (or simply C-means, as it is sometimes known) is a generalization of the "hard" K-means algorithm as described in Bonhomme and Manresa (2015).

In particular, the "hard" K-means (HKM) objective function can be written as

$$L_1(P, \psi) = \int \min_g \|y - \psi_g\|^2 P(dy), \tag{1}$$

where $y \in R^T$ is a vector of outcomes with probability measure P on R^T, $g \in \{1, 2, \ldots, G\}$ indexes groups, and $\psi \in \mathbb{R}^{G \times T}$. ψ_g are known as the "cluster centers". Alternatively, Equation (1) can be rewritten using a weighted sum,

$$L_1(P, \psi) = \int \sum_{g=1}^{G} w_g(y; \psi) \|y - \psi_g\|^2 P(dy), \tag{2}$$

where $w_g = \mathbf{1}\left[\|y - \psi_g\|^2 \leq \|y - \psi_h\|^2 \ \forall\, h \neq g\right]$. The fuzzy C-means (FCM) objective function generalizes the weights w_g in Equation (2) so that they need not be binary. In particular, the objective function is instead

$$J_m(P, \mu, \phi) = \int \sum_{g=1}^{G} \mu_g^m(y; \phi) \|y - \rho_g\|^2 P(dy) \tag{3}$$

where $\mu_g(y; \rho)$ are weights (or a "fuzzy partition" of y), $m > 1$ is a tuning parameter, and $\rho \in R^{G \times T}$. The weights

$$\mu_g(y; \rho) = \left(\sum_{h=1}^{G} \frac{\|y - \rho_g\|^{2/(m-1)}}{\|y - \rho_h\|^{2/(m-1)}} \right)^{-1}, g = 1, \ldots, G, \tag{4}$$

are optimal based on the mean squared error of Equation (3) (Theorem 11.1, Bezdek (1981)), subject to the constraint that the weights $\mu_g(y; \rho)$ sum to unity.

Yang and Yu (1992) show that Equation (3) can be rewritten eliminating the parameter-dependent weights. For fixed m, define $\mu(\rho) = (\mu_1(y; \rho), \ldots, \mu_G(y; \rho))$. Then, a new objective function can be defined as

$$\begin{aligned} L_m(P, \rho) &= J_m(P, \mu(\rho), \rho) \\ &= \int \sum_{h=1}^{G} \left(\sum_{i=1}^{G} \frac{\|y - \rho_h\|^{2/(m-1)}}{\|y - \rho_i\|^{2/(m-1)}} \right) \|y - \rho_g\|^2 P(dy) \\ &= \int \left(\sum_{g=1}^{G} \|y - \rho_g\|^{-2/(m-1)} \right)^{1-m} P(dy). \end{aligned} \tag{5}$$

Significantly, Equation (5) has replaced an objective function with weights linked to the parameters (Equation (3)) with a nonlinear function in $\|y - \rho_g\|$, since the weights themselves are simply a function of $\|y - \rho_g\|$. As far as we know, up to now the representation in Equation (5) has only been used as a device to establish properties of Equation (3), as opposed to an objective function in its own right.

As argued by Yang and Yu (1992), it is clear from $L_m(P, \rho)$ that HKM can be seen as a limiting case of FCM, although it is not strictly nested:

$$L_m(P, \phi) \underset{m \to 1}{\longrightarrow} \int \min_g \|y - \rho_g\|^2 P(dy),$$

as the weights become binary in the limit. In contrast to the discrete assignment of HKM, the tuning parameter, m, allows the researcher to parametrize an appropriate degree of uncertainty over group assignment based on the dataset. In empirically common panel lengths and cross-sectional data, this is an important generalization over the HKM algorithm considered by Bonhomme and Manresa (2015). In this sense, FCM exploits a non-parametrically smoothed version of the HKM weights.

Yang and Yu (1992) prove the equivalence of Equation (3) and Equation (5), establish the existence of a solution, and prove strong consistency of a sample estimator $\hat{\phi}$ for ϕ^*, which minimizes $L_m(P, \rho)$ and $J_m(P, \mu, \rho)$. Yang (1994) extends the analysis of Pollard (1982) to the FCM setting to establish the asymptotic normality of $\hat{\rho}$. In this paper, we consider a more general form of the FCM objective function, just as Bonhomme and Manresa (2015) extend the HKM objective function. Specifically, we are concerned not just with cluster centers ρ (a regression on a group-specific constant), but linear regression more broadly. In particular, we consider an objective function based on

$$\|y - \theta_g x\| \tag{6}$$

where $x \in R^k$ contains random variables and θ_g is $T \times k$. The next section proves several properties of this generalization of the FCM algorithm. We establish an equivalence relationship similar to that between Equations (3) and (5), for the regression counterpart, and prove that a common solution exists. By demonstrating the transformed problem fits into the GMM framework, we harness familiar results (Hansen (1982)) to prove consistency and asymptotic normality of an estimator, $\hat{\theta}$ for θ. We discuss extensions to allow for common coefficients across groups for some set of regressors, as well as an analogue to two-stage least squares.

Before considering these generalizations, in the next section we provide novel results

that analytically characterize the relationship between ψ^*, ρ^*, and the true cluster centers (which we will denote by ζ) in a tractable simple case. These motivate the use of FCM in settings where neither FCM nor HKM will in general recover the true parameters.

2.1 Analytical comparison of HKM and FCM optima

The global optima of both HKM in fixed-T data (ψ^*) and FCM in general (ρ^*) represent pseudo-true parameters. For HKM, this is because the group assignment cannot be consistently estimated in the first stage. For FCM, this is because the objective function does not correspond to the true DGP, since an observation has non-zero weight assigned to multiple groups (whereas the true group assignment is binary). However, we are not aware of any existing analytical results comparing the pseudo-true parameters across estimators. In this section, we characterize this relationship for a tractable simple case. In particular, we consider the problem of estimating the cluster means of two groups, where the data are generated from two Gaussian distributions with distinct means, the same finite variance, and equal mass, where the econometrician observes a single outcome for each observation. We show that even in this simplest of cases, FCM, despite technically being misspecified, can improve on HKM due to having a pseudo-true parameter closer to the truth. Our simulations reported in Section 3, calibrated to the regression problem of our empirical application, bear these results out in a much more complicated setting.

First, we characterize the bias of HKM in this homoskedastic two-Gaussian setting.

Proposition 1. *In the homoskedastic two Gaussian case, ψ_1^*, the HKM global optimum for the lower mean, ζ_1, is given by*

$$\psi_1^* = \Phi\left(-\zeta_1/\sigma\right)\left(\zeta_1 + \frac{-\phi\left(-\zeta_1/\sigma\right)}{\Phi\left(-\zeta_1/\sigma\right)}\right) + \Phi\left(-\zeta_2/\sigma\right)\left(\zeta_2 + \frac{-\phi\left(-\zeta_2/\sigma\right)}{\Phi\left(-\zeta_2/\sigma\right)}\right),$$

and similarly for ζ_2, the higher mean, where σ is the standard deviation of both Gaussian clusters. ζ_1 is negatively biased and ζ_2 is positively biased unless $\sigma \to 0$.

Proposition 1 shows that each of the global minimum of the HKM objective function for each of the cluster means is biased *outwards* (away from zero), with the lower mean being further reduced, and the upper mean being increased. The intuition is that ψ_1^* represents the mean over two truncated normals (portions of the mean ζ_1 and ζ_2 distributions), with the right tail of the distribution with mean ζ_1 truncated and replaced with an equal-mass portion of the ζ_2 distribution that is situated to the left of that tail. In practical terms, this means that HKM overstates cluster heterogeneity in this simple case.

It is much harder to characterize the bias of FCM given the non-linearity of the objective function and the fact that closed-form solutions do not exist for ρ^* except in degenerate cases. However, we offer a result below that situates ρ^* *relative* to ψ^*.

Proposition 2. *In the homoskedastic two Gaussian case, ρ_1^*, the FCM global optimum for the lower mean, satisfies $\rho_1^* \geq \psi_1^*$, and $\rho_2^* \leq \psi_2^*$.*

Proposition 2 shows that the FCM optima are located *inwards* relative to the HKM optima. While it is not possible to make a general statement about their position relative to the true parameters, ζ_1, ζ_2, since the HKM optima are biased outwards, this result means that FCM has the potential to have lower bias than HKM. The distance relative to HKM is increasing in m. Theorem 1 shows that given the correct choice of m, FCM is unbiased in this simple setting.

Theorem 1. *In the homoskedastic two Gaussian case, assuming ρ_1^* and ρ_2^* are unique, there exists some $\tilde{m} \in (1, \infty)$ such that $\rho_1^*(\tilde{m}) = \zeta_1, \rho_2^*(\tilde{m}) = \zeta_2$.*

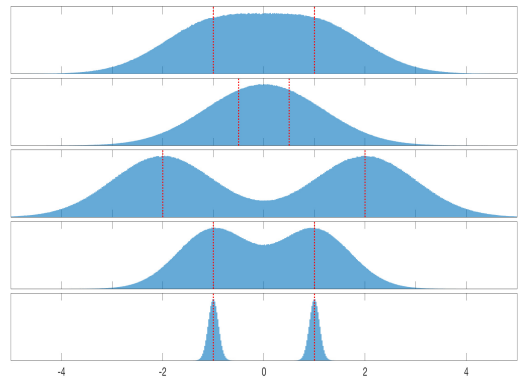
Of course, in practice we do not know $\tilde{m}$, even in this simple case. However, even for non-optimal m, Proposition 2 suggests that the pseudo-true parameters of FCM may be less biased than those of HKM. To demonstrate this, we evaluate numerically the bias of FCM relative to HKM for two values of m (one close to HKM, one in line with empirical practice), as well as $\tilde{m}$, for a number of parameterizations of the two Gaussian setting. Table 1 reports the results, with the right panel displaying the settings visually. In summary, it is clear that some of these examples represent very difficult clustering problems where it is natural to expect HKM to struggle; in these examples, FCM demonstrates smaller bias due to its ability to accommodate uncertainty over group membership. However, HKM does demonstrate smaller bias when the data are better separated (row 3). The distance to HKM is increasing in m, and, depending on the parameterization, FCM may be biased outwards, like HKM, or biased inwards. These results show that HKM and FCM recover different pseudo-true parameters for the cluster means, and that in difficult clustering problems with worse-separated data, the FCM pseudo-true parameters may exhibit lower bias in an analytically tractable setting. Of course, given the results of Theorem 1, when $m = \tilde{m}$, the bias of FCM is zero, as displayed in the final column. While such closed-form results cannot be extended to the more general setting of our paper, they motivate the use of FCM in settings where T is small, so HKM is known to not recover the true parameters.

Figure 1: HKM and FCM performance in a simple 2-group model

(a) Numerical values of ψ_2^* and ρ_2^*

σ^2	ζ_2	HKM		FCM	
(I)	(II)	(III)	(IV)	(V)	(VI)
			$m = 1.1$	$m = 1.8$	$m = \tilde{m}$
1.0	1.0	1.17	1.17	1.04	1.00
1.0	0.5	0.90	0.90	0.77	0.50
1.0	2.0	2.02	2.02	1.92	2.00
0.5	1.0	1.14	1.05	0.96	1.00
.01	1.0	1.00	1.00	1.00	1.00

(b) Distribution of data



Notes: Table 1 reports the HKM and FCM estimates (for both $m = 1.1$ and $m = 1.8$) in columns III-V. Each row corresponds to a different empirical setting, with the variance and higher mean listed in columns I-II. Column VI reports the estimates for the $\tilde{m}$ which delivers an unbiased FCM estimate. Appendix C.1 displays ρ_2^* as a function of m. Based on these results, $\tilde{m}$ is equal 1.9, 2.4, 1.5, 1.7 and 1.01 respectively. Figure 1 depicts visually the distribution of the data in each of the empirical setting studied. The first row in Table 1 corresponds to the top panel in Figure 1 and so on.

2.2 Properties of fuzzy C-means

We begin by generalizing the FCM objective function for cluster centers to a regression problem. Consider the model

$$y_i = \sum_{g=1}^{G^*} \mathbf{1}\left[i \in g\right] \theta_g x_i + \varepsilon_i, \quad i = 1, \ldots, N \tag{7}$$

where $y_i \in \mathbb{R}^T$, $x_i \in \mathbb{R}^k$, (y_i, x_i) are i.i.d. and $E\left[\varepsilon_i \mid x_i\right] = 0$, according to the probability measure Π on y, x (denoting the conditional for y by $\Pi_{y|x}$ and the marginal for x by Π_x), and θ_g is a $T \times k$ matrix. Equation (7) postulates that the outcomes, y, are generated linearly from x, with the parameters depending on observation i's group membership, captured by the indicators $1\left[i \in g\right]$. However, the group assignments $1\left[i \in g\right]$ are unknown, and in general cannot be recovered with certainty.[9] The natural FCM version of Equation (7) for G groups is given by J_m^{reg}:

$$J_m^{reg}\left(\Pi, \mu^{reg}, \theta\right) = \int \int \sum_{g=1}^{G} \left(\mu_g^{reg}\left(y \mid x; \theta\right)\right)^m \left\|y - \theta_g x\right\|^2 \Pi_{y|x}\left(dy \mid x\right) \Pi\left(dx\right), \tag{8}$$

[9]For the moment, we take G as given, but return to the choice of G in Section 2.3.

where $\theta \in \Theta \subset R^{G \times T \times k}$, and

$$\mu_g^{reg}(y \mid x; \theta) = \left(\sum_{h=1}^{G} \frac{\|y - \theta_g x\|^{2/(m-1)}}{\|y - \theta_h x\|^{2/(m-1)}} \right)^{-1}, g = 1, \ldots, G.$$

Denote $\mu^{reg}(\theta) = \left(\mu_1^{reg}(y \mid x; \theta), \ldots, \mu_G^{reg}(y \mid x; \theta)\right)$. The objective function in Equation (8) involves parameter-dependent weights. Existing implementations and convergence results for FCM (based on group means only) suggest that an iterative procedure, updating $\mu\left(\hat{\theta}^{(r-1)}\right)$ based on a previous estimate $\hat{\theta}^{(r-1)}$ and using these weights to estimate and update $\hat{\theta}^{(r)}$, will converge and may in fact be consistent for θ. However, both from theoretical and computational standpoints, it is desirable to work with a more compact, weight-free representation. Theorem 2 generalizes the equivalence result of Yang and Yu (1992) to the regression problem of Equation (8).

Assumption 1. *Observations (y_i, x_i) are generated according to (7), jointly i.i.d. with probability measure Π, G is finite, and $E[\varepsilon_i \mid x_i] = 0$.*

Theorem 2. *(Equivalence) Under Assumption 1, $J_m^{reg}(\Pi, \mu^{reg}, \theta) = L_m^{reg}(\Pi, \theta)$, where*

$$L_m^{reg}(\Pi, \theta) = \int \int \left(\sum_{g=1}^{G} \|y - \theta_g x\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx); \tag{9}$$

a minimizer $\theta^* \in \Theta$ of $L_m^{reg}(\Pi, \theta)$ is also a minimizer of $J_m^{reg}(\Pi, \mu^{reg}, \theta)$ over Θ and weights μ^{reg}.

Theorem 2 demonstrates that we can turn our attention from the objective function $J_m^{reg}(\Pi, \mu^{reg}, \theta)$ to the simpler formulation in $L_m^{reg}(\Pi, \theta)$, without weights. Under additional regularity conditions, we now establish the existence of a solution to the FCM problem in Equation (8).

Assumption 2. *Assume*

1. *The second moments of y and x are finite under Π:*

$$\int \int \|y\|^2 \Pi_{y|x}(dy \mid x) \Pi(dx) < \infty,$$

$$\int \|x\|^2 \Pi(dx) < \infty,$$

$$\int \int \|y\| \|x\| \Pi_{y|x}(dy \mid x) \Pi(dx) < \infty,$$

2. *Additionally, the x are not collinear,*

$$rank\left(\int xx'\Pi\left(dx\right)\right)=k.$$

3. Θ *is compact.*

Theorem 3. *(Existence) If Assumptions 1-2 hold, then for any $g = 1, 2, \ldots$, there exists a solution θ^* such that*

$$L_m^{reg}\left(\Pi,\theta^*\right)=\inf_{\theta}L_m^{reg}\left(\Pi,\theta\right).$$

The relationship between θ^* and the true parameter, θ_0, generating the data in Equation (7) merits further discussion. θ^* is a *pseudo-true* parameter, which will not be in general be equivalent to θ_0, since it optimizes an objective function that is not perfectly aligned with the the true data generating process (DGP). This is because the objective function puts non-zero weight on observations being members of groups other than their (unknown) true group. This is also the case for HKM in Bonhomme and Manresa (2015) for fixed T; only a pseudo-true parameter can be recovered, since group membership cannot be consistently estimated, even though the objective function appropriately represents the DGP. In general, the closer m is to unity, the closer the FCM objective function corresponds to the true DGP, but the less uncertainty can be accommodated. Under two special limiting cases, however, $\theta^* = \theta_0$. First, as the degree of separation of the groups diverges, $\sum_{h=1}^{G}\frac{\left\|y-\theta_{g_0}x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}\to 1$,where g_0 is the true group, so the FCM objective function L_m^{reg} converges to its HKM counterpart. Second, as $m \to 1$, under additional weak dependence assumptions (like those in Bonhomme and Manresa (2015)), the weight on the true group will likewise converge to unity asymptotically in T. While neither of these cases is a plausible description of most macroeconomic datasets, we present evidence in Section 3.2 that the pseudo-true parameters may still be very close to the true parameters in practice.

Up to this point, our results generalize existing FCM results from Yang and Yu (1992) to the regression problem in Equation (6) in order to establish that a solution to the modified FCM problem exists and can be obtained from Equation (9). Previously, the asymptotic properties of both HKM and FCM problems have proven quite difficult to establish, requiring extensive technical arguments (e.g., Pollard (1981, 1982), Yang and Yu (1992), Yang (1994), Bonhomme and Manresa (2015),Bonhomme and Manresa (2015)). However, Theorem 4 shows that the solution to the FCM problem has a familiar form.

Theorem 4. *(Moments) The solution θ^* satisfies the moment equations*

$$E\left[\left(\sum_{h=1}^{G}\frac{\left\|y_i-\theta_g x_i\right\|^{2/(m-1)}}{\left\|y_i-\theta_h x_i\right\|^{2/(m-1)}}\right)^{-m}\left(y_{it}-\theta_{g,(t)}x_i\right)x_i\right]=0 \text{ for } g=1,\ldots,G \text{ and } t=1,\ldots,T, \tag{10}$$

where t indexes dimensions of y_i and (t) rows of θ_g; FCM is a GMM problem.

Theorem 4 shows that FCM constitutes a standard generalized method of moments (GMM) problem (Hansen (1982)). This has two important implications. First, reframing FCM as a GMM problem allows the asymptotic properties of estimators $\hat{\theta}$ to be derived using standard theory. Second, existing implementations of FCM have focused on iterative procedures based on Equation (8). In this formulation, the weights $\mu^{reg}(\theta)$ must be simultaneously determined, as in weighted least squares, and then used to re-estimate θ, and so on, until convergence. However, rather than just being a convenient theoretical device, the representation in Equation (9) facilitates nonlinear optimization via a single step procedure, with the familiar apparatus of GMM. We return to this second point in Section 3.1.

Additionally, the moment equations in Theorem (4) can easily accommodate regressors with common coefficients across groups, $\theta_{g,tk}=\theta_{h,tk}$, or across dimensions of y_i, $\theta_{g,tk}=\theta_{h,sk}$.[10] In the former case, it is straightforward to show that the corresponding moment condition is

$$E\left[\left(\sum_{h=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2/(m-1)}\right)^{-m}\sum_{g=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2m/(m-1)}\left(y_{it}-\theta_{g,(t)}x_i\right)x_{ik}\right]=0$$

and in the latter,

$$E\left[\left(\sum_{h=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2/(m-1)}\right)^{-m}\sum_{t=1}^{T}\sum_{g=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2m/(m-1)}\left(y_{it}-\theta_{g,(t)}x_i\right)x_{ik}\right]=0.$$

The moment conditions in Equation (10) have natural sample counterparts that can be used to define the estimator $\hat{\theta}$.

[10] To trace a direct link with our baseline specification forthcoming in Equation (14), the parameter set θ encompasses both group-specific constants α_g and MPCs, as well as the coefficients on the common covariates W.

Definition 1. Let the estimator $\hat{\theta}$ be the solution to

$$S_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \eta(\theta, y_i, x_i)' \sum_{i=1}^{N} \eta(\theta, y_i, x_i), \tag{11}$$

where

$$\eta(\theta, y_i, x_i) = \left[\left(\sum_{h=1}^{G} \frac{\|y_i - \theta_g x_i\|^{2/(m-1)}}{\|y_i - \theta_h x_i\|^{2/(m-1)}}\right)^{-m} \left(y_{it} - \theta_{g,(t)} x_i\right) x_i\right] = 0$$
$$\text{for } g = 1, \ldots, G \text{ and } t = 1, \ldots, T,$$

the $(G \times T \times k) \times 1$ vector-valued moment function.

Assumption 3. *θ^* is the unique solution to $E[\eta(\theta, y_i, x_i)] = 0$ (up to ordering of the groups).*

Assumption 3 is an identification condition. In clustering models, if identification holds, it does so up to a labeling of the groups g, which can always be permuted. The FCM literature (and HKM when T is fixed, e.g., Pollard (1981, 1982)) always assumes the uniqueness of θ^*. While it appears intractable to characterize primitive conditions under which identification holds in the non-linear form of Equation (9), it is closely linked to the OLS identification. With known membership, Assumptions 1-2 suffice for OLS to uniquely identify the true parameters, and for relatively small m, the objective (9) is a perturbation around the OLS objective. In a formal sense, identification in clustering models when group membership cannot be recovered remains a topic for future work. Theorem 5 establishes consistency.

Theorem 5. *(Consistency) Under Assumptions 1-3, $\hat{\theta} \xrightarrow{p} \theta^*$ as $N \to \infty$.*

With additional assumptions, the asymptotic distribution of $\hat{\theta}$ can be characterized.

Assumption 4. *Additionally,*

1. *θ^* is in the interior of Θ,*
2. *$H = E\left[\frac{\partial \eta(\theta, y_i, x_i)}{\partial \theta'}\right]$ is full rank,*
3. *$E\left[\sup_{\theta \in \mathcal{N}} \left\|\frac{\partial \eta(\theta, y_i, x_i)}{\partial \theta'}\right\|\right] < \infty$ in a neighborhood $\mathcal{N}$ of θ^*,*
4. *$E\left[\eta(\theta^*, y_{i,}, x_i)\, \eta(\theta^*, y_{i,}, x_i)'\right]$ is positive definite.*

These are largely technical conditions. Theorem 6 gives the limiting distribution of $\hat{\theta}$.

Theorem 6. *(Asymptotic Normality) Under Assumptions 1 - 4,*

$$\sqrt{N}\left(\hat{\theta}-\theta^{*}\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1} V H^{-1}\right),$$

where

$$V = E\left[\eta\left(\theta, y_i, x_i\right) \eta\left(\theta, y_i, x_i\right)'\right],$$

and H is the Hessian of Equation (9).

We provide explicit expressions for H in Appendix Section B, including the case of common coefficients across groups. This result (as well as the underlying assumptions) closely parallels that of Yang (1994), who establishes asymptotic normality for the simpler cluster centers case.

Two important distinctions relative to HKM remain to be discussed. First, at no point have we relied on an assumption that the data are well-separated, unlike Bonhomme and Manresa (2015). They require this assumption so that the true group membership function can be consistently estimated. However, in FCM there is no need to estimate such a function; ultimately, L_m^{reg} does not even have group-specific weights to be estimated. This means that consistency for the pseudo-true θ^* and asymptotic normality of the estimates hold regardless of whether the groups are in fact well-separated. Nevertheless, it is the case that the pseudo-true FCM parameters converge to the true parameters as separation increases. Bonhomme and Manresa (2015) identify methods for inference that are robust to a lack of group separation as an important question for future work. We suggest that FCM presents such an option. In Section 3.2, we present evidence that FCM performs very well in empirically calibrated simulations, while the empirical noise is such that separation of the groups is doubtful and HKM performs poorly.

Second, we have at no point made assumptions on the relationship between G, the econometrician's number of groups, and G_0, the true number of groups. This has been possible since our results are relative to θ^*, a pseudo-true parameter, which is defined with respect to G; this is also the case for the discussion of fixed-T inference in Bonhomme and Manresa (2015). However, if $G \neq G_0$, then there is reason to doubt the closeness of the relationship between θ^* and θ_0, so selection of G remains an important question. Unfortunately, the vast majority of methods to select G are derived under large N, T asymptotics, which do not apply in our setting. For this reason, we discuss a flexible non-parametric approach for selecting optimal G in Section 2.3.

Up to now, we have maintained the standard OLS assumptions, with strictly exogenous regressors x, such that $E\left[\varepsilon_i \mid x_i\right] = 0$. However, in some regression contexts, the regressor of interest is potentially endogenous. A standard solution in the regression con-

text is to use an instrumental variable(s), z, when available. If relevance ($cov\left(x_i, z_i\right) \neq 0$) and exogeneity ($E\left[\varepsilon_i z_i\right] = 0$) conditions are satisfied, then the coefficient on the endogenous x can be recovered via two-stage least squares (TSLS). Since FCM only permits the recovery of pseudo-true parameters for the true model in Equation (7) even under strict exogeneity, the motivation for IV is not as direct as in standard models, because it does not necessarily allow the true parameters to be recovered. However, IV estimates do have a similar interpretation as the marginal effect of *exogenous* variation in x on y. Moreover, as with OLS, if either the degree of separation diverges or m tends to unity and T to infinity while additional weak dependence assumptions are maintained, TSLS with a valid instrument will recover the true coefficient on an endogenous regressor. We consider such a TSLS estimator in Section 5.[11] In particular, for an endogenous regressor x^e and additional controls ω, we estimate the first stage-regression

$$x_i^e = \gamma z_i + \tau \omega_i + u_i,$$

via OLS, and then generate $\tilde{x}_i^e = \gamma z_i + \tau \omega_i$. We define $\tilde{x} = \begin{pmatrix} \tilde{x}^e & \omega \end{pmatrix}'$, and input these predicted values in Equation (11) to estimate the second stage. In Appendix Section B, we extend all of the theoretical results presented above to this alternative problem, establishing the equivalence of the representations, existence of a solution in population, consistency of a sample estimator for that solution, and asymptotic normality of the estimator. We also provide analytical expressions for the asymptotic variance. In keeping with our empirical problem, we focus on the case of a single endogenous regressor with a single instrument, but the results can be easily extended.

Our theoretical development has focused on linear regression models. However, there is no reason that FCM could not similarly be applied to more general problems. In a nonlinear regression setting, for example, the residual from the linear regression function, $y_i - \theta_g x_i$, would simply be replaced by $y_i - f\left(\theta_g, x_i\right)$, for some nonlinear function $f\left(\cdot\right)$ (which, in principle, could itself depend on g). The theoretical results established here would surely extend, following suitable modifications of the assumptions to accommodate the behavior of the function in question. To this extent, the FCM methodology is highly flexible, and has the potential to be employed in a wide range of contexts exhibiting potentially heterogeneous relationships in economics.

[11] To foreshadow what will come in our empirical context, we use this TSLS framework to instrument for the rebate value - which is potentially endogenous - with the rebate receipt.

2.3 Choosing the number of groups G

For a given G, it is easy to apply the fuzzy C-means algorithm as described above. However, G is unknown. To choose G, we extend the "gap statistic" from Tibshirani et al. (2001) to the regression setting. In the Tibshirani et al. (2001) setting, the researcher has data on some characteristic for observation i, y_i, for each $i \in N$, where N is the number of observations. Define the residual sum of squares $W_{ss}(G)$ under G groups as:

$$W_{ss}(G) = \sum_{g \in G} \frac{1}{2N_g} \sum_{i',i \in g} d_{i'i}$$

where $d_{i'i}$ is the Euclidean distance between observations i and i' and N_g is the number of observations belonging to cluster g. The gap statistic identifies the number of clusters by comparing the average within group sum of squares under G groups ($W_{ss}(G)$) to the *expected* within group sum of squares under G groups under the assumption that there is no clustering present in the data (the "reference distribution"). Formally, the gap statistic is:

$$\text{Gap}(G) = E\left[log\left(W_{SS}(G)\right)\right] - log\left(W_{SS}(G)\right) \tag{12}$$

The expected within group sum of squares is calculated on many samples of simulated data generated under the assumption that there is no clustering present in the data. This is achieved by sampling the outcomes from a uniform distribution on an interval $[a, b]$ where a and b are chosen as the maximum and minimum observed values of the outcomes in the sample. Ideally, G is chosen such that the gap statistic is maximized. The basic idea is to normalize the within group sum of squares (WSS) curve by what one would expect to get regarding the WSS in a sample in which no clusters are present, and any improvement in the WSS is simply due to fitting noise. In practice, Tibshirani et al. (2001) find that the gap statistic may display local maxima when the data are not well-separated or sub-clusters are present, and it is advisable to inspect the "gap curve" as opposed to mechanically choosing its maximum.

The objective W_{SS} above is tailored to the cluster mean setting of Tibshirani et al. (2001). We opt to use our regression objective as a natural alternative: the weighted sum of squared residuals, Equation (8). The selection procedure for G then proceeds as follows:

1. Run a homogeneous regression (one that assumes no group heterogeneity). Call the residuals from the regression $\hat{\epsilon}_i$.

2. Generate B samples of simulated outcomes for each observation using the homogeneous coefficients, with an error that is uniformly distributed from the minimum to the maximum of $\hat{\epsilon}_i$.

3. Run the fuzzy C-means algorithm described in Section 1 on each of the B samples and for each $G \in \bar{G}$, where $\bar{G}$ is some upper bound. For each sample and each $G \in \bar{G}$, compute the weighted sum of squared residuals, Equation (8).

4. Run the fuzzy C-means algorithm on the actual data and compute the same object.

5. Choose $G \leq \bar{G}$ corresponding to the maximum value of (12) that is statistically significantly greater than the value for all $g \in \{1, ..., G-1\}$ and displays a positive gradient.

The final point operationalizes the advice of Tibshirani et al. (2001) in not simply choosing the maximum of the gap statistic. In particular, the requirement that the chosen G represents a *statistically significant* increase over all previous estimates through $G-1$ avoids spuriously increasing G when doing so does not offer a significantly better representation of the data. We introduce the positive gradient requirement to ensure we identify important sub-clusters and to address the possible non-monotonicity noted by Tibshirani et al. (2001).[12]

3 FCM and HKM performance in simulations

In this section, we discuss the advantages of the FCM methodology in practice. First, we consider computational advantages in the context of realistically-structured cross-sectional economic datasets. We then present the results of simulation studies, in particular comparing FCM to HKM approaches, as well as the performance of the gap statistic.

3.1 Computational tractability

Our approach is computationally tractable and entails a sizable improvement compared to alternative techniques. First, the equivalence shown in Theorem 2 implies a reduction in the computational time to solve the algorithm compared to the existing iterative

[12]This is particularly important in our empirical setting, where we group both on intercepts and MPCs, but are interested only in the latter, since it is conceivable that the first level of clustering might only represent level heterogeneity in consumption changes. We additionally find that - in simulations in which the data are not well-separated - without this requirement the gap statistic has a tendency to erroneously favor homogeneous models.

procedure, while improving precision. More generally, the FCM approach is considerably faster than HKM. In the simulations outlined in the following subsection we consider two versions of the algorithm proposed by Bonhomme and Manresa (2015). Appendix C.2 discusses the algorithms and presents some details on the computational performance. We show next that FCM also outperforms even an enhanced version of the baseline HKM algorithm, conditional on computational feasibility described in the appendix.

3.2 Simulation performance

To explore the comparative performance of FCM and HKM in practice, we conduct a Monte Carlo study based on our empirical application. We calibrate our simulations to our empirical baseline specification, which is forthcoming in Equation (14). Following Bonhomme and Manresa (2015), we generate data according to this model, for total expenditures (by fixing the values of the regressors to those from the true dataset, and generating Gaussian errors $\tilde{\epsilon}_i$ with a calibrated variance or resampling from the empirical distribution). For each observation, we assign a true group based on the modal weight recovered in our empirical study.

We consider four different estimation approaches. First, we estimate via FCM with both $m = 1.8$ and $m = 1.1$. The choice of $m = 1.8$ mirrors suggestions in the literature; we additionally assess $m = 1.1$ as an implementation close to the limiting case of HKM, as discussed above.[13] Next, we consider HKM Algorithm 1 from Bonhomme and Manresa (2015), as well as a version of HKM Algorithm 2 from the same paper.[14]

Performance for $G = 5$

We begin by assessing the performance with a moderate number of groups, $G = 5$. This corresponds to the specification selected by the gap statistic in our empirical setting. We first calibrate the standard deviation of $\tilde{\epsilon}_i$ to 583, the empirical value. We then generate errors by resampling from the empirical distribution within each group.

Table 1 reports the results with normally-distributed errors. Panel 1 displays the mean

[13] The literature generally suggests $1.5 \leq m \leq 2.5$ (e.g., Bezdek et al. (1984), Pal and Bezdek (1995), Yu, Jian et al. (2004), Wu (2012)). While optimality results are not generally available (theoretically or numerically), Yu, Jian et al. (2004) derive a data-dependent theoretical upper bound for m, below which the simple sample mean is not recovered (in the cluster means setting); Wu (2012) shows that in some datasets this bound can be as low as 1.77. Results in Torra (2015) generally favor $1.5 \leq m \leq 2$. Following our reading of these existing results, and the confirmatory results we obtain in this section, we adopt $m = 1.8$ in our empirical study.

[14] In every simulation, we order groups by choosing the group order which minimizes the sum of the squared error (summing over each FE and MPC) between the truth and the estimators.

	Truth	FCM		HKM	
		$m = 1.8$	$m = 1.1$	Algo. 1	Algo. 1.5
Point Estimates	0.651	0.653	0.653	0.650	0.650
	0.423	0.419	0.422	0.423	0.423
	0.245	0.248	0.245	0.246	0.246
	0.516	0.517	0.515	0.516	0.516
	0.289	0.293	0.293	0.289	0.289
RMSE		0.053	0.053	0.055	0.055
		0.026	0.025	0.025	0.025
		0.017	0.016	0.017	0.016
		0.023	0.023	0.025	0.024
		0.054	0.054	0.053	0.052
Rejection Rates		0.042	0.044	0.046	0.046
		0.056	0.052	0.012	0.012
		0.058	0.052	0.030	0.030
		0.032	0.036	0.012	0.012
		0.058	0.058	0.060	0.056
Share Misclassified		0.000	0.000	0.000	0.000

Table 1: Simulation, Gaussian errors, empirical noise, $S = 500$

point estimates for each MPC against the true values used to generate the data for each of the four estimators. For this parsimonious model, all estimators deliver essentially the true parameter values on average. The second panel reports the RMSE of each estimate, which are all quite small and very similar. Additionally, misclassification (based on the modal weight for FCM) is essentially zero across approaches. The third panel additionally reports rejection rates for nominal 5% tests of each true MPC for each estimator using the inference results derived in this paper.[15] These rates, close to 5%, demonstrate the inference methods proposed are well-sized, and suitable for empirical use in datasets with our sample size. In summary, all estimators perform well for this specification.

Table 2 reports parameter estimates with errors sampled from the empirical distribution, which results in data that are much less well-separated. The results are more varied than for the Gaussian errors. No method precisely estimates the highest MPC correctly, with both FCM approaches actually missing by more than HKM. However, for the remaining MPCs, FCM with $m = 1.8$, corresponding to the highest level of smoothing, produces the best results. With the exception of the second MPC, it recovers the true values very precisely, while all other approaches miss by some distance. The RMSE tells a

[15] In Section B.3 of the Appendix, we extend small T inference results from Bonhomme and Manresa (2015) to the case of heterogeneous slope coefficients that we consider.

	Truth	FCM		HKM	
		$m = 1.8$	$m = 1.1$	Algo. 1	Algo. 1.5
Point Estimates	0.651	0.712	0.558	0.691	0.630
	0.423	0.341	0.201	0.292	0.252
	0.245	0.238	0.156	0.187	0.170
	0.516	0.529	0.376	0.395	0.386
	0.289	0.334	0.217	0.226	0.207
RMSE		0.792	0.783	0.847	0.791
		0.317	0.459	0.399	0.429
		0.106	0.150	0.168	0.144
		0.458	0.414	0.494	0.410
		0.468	0.474	0.463	0.468
Share Misclassified		0.053	0.048	0.064	0.055

Table 2: Simulation, empirical errors, $S = 500$

similar story.[16] Misclassification rates are roughly consistent across approaches at 5%.

Performance for $G = 10$

We now turn to the $G = 10$ specification with empirical noise (261). As more groups are introduced, clusters naturally become less well-separated, posing a sterner challenge to the estimators. The first panel of Table 3 reports the mean estimates. Remarkably, even in this very poorly-separated data, FCM with $m = 1.8$ continues to estimate the true parameters quite closely. The additional smoothing introduced for the higher value of m improves the estimator's ability to overcome uncertainty in assignment; $m = 1.1$ now shows weaker performance, substantially misestimating several MPCs. However, FCM with $m = 1.1$ still displays an advantage over both HKM estimators, which misestimate most of the MPCs. These results are further supported by the RMSE, reported in the second panel; FCM with $m = 1.8$ recovers the true parameters quite precisely, despite the challenges posed by the additional clusters and noisy data. The misclassification rates reported below again demonstrate this advantage.

The third panel additionally reports rejection rates for nominal 5% tests. These remain below 10% across MPCs for $m = 1.8$, with most close to 5%, but performance is naturally weaker for the other estimators due to the bias in the estimates themselves.

In appendix C.3 we show that HKM mis-estimation of some MPCs is not confined to a small share of observations. In contrast, our benchmark FCM model does a remarkably better job at matching the empirical CDF of the MPC distribution from the DGP. More-

[16] A possible exception is the fourth MPC, where the $m = 1.8$ result is affected by outliers.

	Truth	FCM		HKM	
		$m = 1.8$	$m = 1.1$	Algo. 1	Algo. 1.5
Point Estimates	0.844	0.784	0.668	-4.129	0.069
	0.986	1.032	2.136	12.602	7.731
	0.795	0.797	1.049	-1.225	1.177
	0.646	0.646	0.499	-0.466	0.551
	0.496	0.495	0.456	0.483	0.427
	0.468	0.477	0.460	0.667	0.382
	0.496	0.488	0.401	0.440	0.813
	0.268	0.269	0.354	0.365	0.351
	0.340	0.344	0.443	0.030	0.130
	0.257	0.263	0.505	0.433	0.483
RMSE		0.676	5.932	10.110	8.869
		0.771	17.218	20.519	16.644
		0.029	3.829	6.737	5.222
		0.023	2.530	4.298	3.323
		0.094	1.599	1.890	1.765
		0.129	3.547	2.246	2.258
		0.174	1.769	2.143	2.111
		0.023	0.389	0.484	0.649
		0.048	0.910	1.344	1.391
		0.056	0.421	0.410	0.471
Rejection Rates		0.088	0.758	0.982	0.962
		0.106	0.928	0.998	0.984
		0.074	0.904	0.954	0.914
		0.086	0.892	0.938	0.876
		0.066	0.882	0.770	0.704
		0.058	0.880	0.744	0.700
		0.060	0.754	0.750	0.674
		0.054	0.456	0.554	0.458
		0.070	0.610	0.720	0.902
		0.058	0.508	0.536	0.632
Share Misclassified		0.007	0.409	0.775	0.743

Table 3: Simulation, Gaussian errors, empirical noise, $G^* = 10$, $S = 500$

over, we explore whether HKM can be enhanced by increasing the tuning parameters governing computational time. While performance slightly improves, it is still weaker than our benchmark model.

Gap statistic performance

We also assess the performance of the gap statistic in selecting the correct number of groups. For this simulation, we draw 50 samples using as true structure $G = 5$ and non-parametrically resampling from the empirical distribution of errors for each group to preserve its properties. We find that in all 50 samples, the gap statistic correctly selects $G = 5$ as the preferred specification. This evidence corroborates the original finding of Tibshirani et al. (2001) that the gap statistic performs well for mean clustering problems in our regression context. As an additional check on our methodology for selecting G, we consider the Partition Coefficient and Partition Entropy measures proposed by Bezdek (1981). These measures are not designed to select an "optimal" grouping, but rather indicate specifications that are supported by the data. In the data, both support the $G = 5$ specification, increasing in value from $G = 4$. Additionally, we find that both measures strongly favor the $G = 5$ specification in all 50 simulated samples.

In further related simulations, we also consider the consequences of mis-specifying G. We estimate a specification with 5 groups when the true structure has 10 groups, and vice versa. We find that quantiles of the distribution of coefficients recovered when G is underspecified matches the true distribution quite well. The performance for over-specified G is worse, since the additional coefficients spuriously fit noise in the tails of the data. These results (and the computational challenges of richer models) suggest it may be desirable to err towards models with fewer groups.

Discussion

The simulation results show that FCM, particularly with higher values of m, can exhibit advantages over HKM when the data is less well-separated, which in practice may coincide with models with more groups or, in our data, empirically-distributed (non-Gaussian) noise. Interestingly, these results show that even though FCM is inherently misspecified with respect to Equation (7), while HKM is well-specified, its performance is not necessarily inhibited. In spite of the smoothing introduced by the weights relative to the true objective function, both FCM and HKM recover the same parameters in lower noise environments, and FCM may in fact remain closer to the truth than HKM in higher noise or weakly-separated environments. We conjecture that this difference in perfor-

mance is driven by the fact that observations that HKM would misclassify are not given binary assignments by FCM, rather having some impact on estimates for their true group, and a reduced impact on parameters of the incorrect group that HKM would assign. This is supported by the fact that the weakened performance of HKM that we observe for $G = 10$ is associated with an increased probability of misclassification. This interpretation is consistent with the analytical results reported in Section 2.1 for the Gaussian cluster means case. The accommodation of uncertainty in group assignment via continuous weights improves the performance of the FCM estimator.

4 Empirical methodology

We now apply our estimator to investigate heterogeneity in the marginal propensity to consume, focusing our analysis on the 2008 Economic Stimulus Act (ESA), as in Parker et al. (2013). Between April and July of 2008, $100 billion in tax rebates was sent to approximately 130 million US tax filers.[17] The timing of the rebate receipt was determined by the last two digits of the recipient's Social Security Number (SSN), making the timing of receipt random. As in Parker et al. (2013), we also exploit the randomized timing of the rebate receipt, but instead estimate heterogeneous (and unobserved) propensities rather than a homogeneous marginal propensity to consume. Our data come from the Consumer Expenditure Survey (CEX), which contains comprehensive and detailed measures of household-level consumption expenditures. The 2008 CEX wave also includes supplemental questions on the ESA, including the amount of each stimulus payment received. While CEX expenditures are reported at the quarterly frequency, new households enter the survey at each month, making the frequency of our data monthly. Since we depart from Parker et al. (2013) by allowing for treatment heterogeneity, we present their homogeneous specification first as a useful benchmark, introducing our refinements thereafter.

4.1 Homogeneous MPC

Parker et al. (2013) consider the following specification:

$$\Delta C_j = \beta' W_j + \theta R_j + \alpha + \epsilon_j \tag{13}$$

where ΔC_j is the first difference of consumption expenditure of household i in quarter

[17] We defer to Parker et al. (2013) and Sahm et al. (2010) for an exhaustive discussion of the Economic Stimulus Act.

t.[18] W_j is a set of controls including month dummies aimed at absorbing common time effects such as aggregate shocks, as well as seasonal factors.[19] The independent variable of interest is R_j, which denotes the amount of the tax rebate received by each household. θ is then interpreted as the causal effect of the rebate on expenditures, where identification is achieved by comparing expenditure changes of households that received the rebate in a certain period to expenditure changes of households that did not receive the rebate in the same period.[20]

4.2 Heterogeneous MPCs

We depart from the homogeneous specification in Equation (13) and allow for heterogeneity in the expenditure responses to the tax rebate across households. In particular, we follow the structure of Equation (7) and augment Parker et al. (2013)'s specification as follows:

$$\Delta C_j = \beta' W_j + \sum_{g \in G} \left(\theta_g \mathbf{1}\left[j \in g\right] R_j + \alpha_g \mathbf{1}\left[j \in g\right] \right) + \epsilon_j \tag{14}$$

That is, we assume that heterogeneity in responses to the rebate can be summarized with G groups, characterized by the vector of coefficients $\{\alpha_g, \theta_g\}$. We include the group-specific intercepts α_g to correctly interpret θ_g as a marginal propensity to consume. For example, since we cannot control for changes in income, without the group-specific level effects, MPC heterogeneity might be biased by heterogeneity in income changes unrelated to the tax rebate.[21] $\mathbf{1}\left[j \in g\right]$ is an indicator that takes a value of 1 if household i in period t belongs to a certain group $g \in G$. Our object of interest is θ_g, which describes MPC het-

[18] To maintain consistent notation throughout the paper, we refer to j as the (i, t) combination. We wish to emphasize that while we have information on the same households i in different periods t, identification is not obtained by comparing individual responses over time. We do not exploit any limited panel structure, except to construct consumption changes for the left-hand-side variable. We return to this point below.

[19] In Parker et al. (2013), the other controls are age, change in number of adults in the household, and change in the number of children in the household. The controls we will use are the same, but additionally include squared age.

[20] As discussed by Kaplan and Violante (2014), θ may not correctly measure the marginal propensity to consume out of a transitory income shock, but is instead better thought of as a "rebate coefficient". This is because the control group of non-recipients in period t is made of three groups: (i) households that never receive the rebate, (ii) households that did not receive a rebate yet, but may anticipate receiving the rebate in the future, and (iii) households that have already received the rebate. The second group might display a positive MPC out of news of the rebate, biasing the estimated rebate coefficient θ downward. Similarly, the third group might also have a positive lagged MPC out of the rebate, further contributing to a downward bias. We address these issues in Appendix Section D.3.

[21] The CEX has information on current income during the first interview, but not thereafter, so we cannot construct measures of income changes for each quarter.

erogeneity, while $\mathbf{1}[j \in g]$ tells us the group membership of each household. The vector of coefficients, combined with $1[j \in g]$, gives an approximation of the MPC distribution. Section B.4 of the Appendix discusses the distinction between recovering MPCs based on a parsimonious specification as in Equation (14) and subsequently investigating their relationship with additional covariates, as compared to including additional covariates in the regression itself.

5 Results

We apply our FCM approach to the rebate experiment, estimating Equation (14). Our findings highlight a considerable degree of MPC heterogeneity whose extent varies depending on the consumption category considered. We first show the distribution of marginal propensities to consume out of total expenditures and illustrate how our results are robust to different specifications and sample selection procedures. We then investigate how the MPC distribution changes as we consider nondurable and durable goods as the dependent variables. Importantly, our approach also allows us to directly test whether households display similar propensities for different consumption goods, or instead substitute across expenditure types when they receive a transitory income shock such as a tax rebate. Finally, we explore which observable household characteristics are correlated with the estimated marginal propensities to consume.

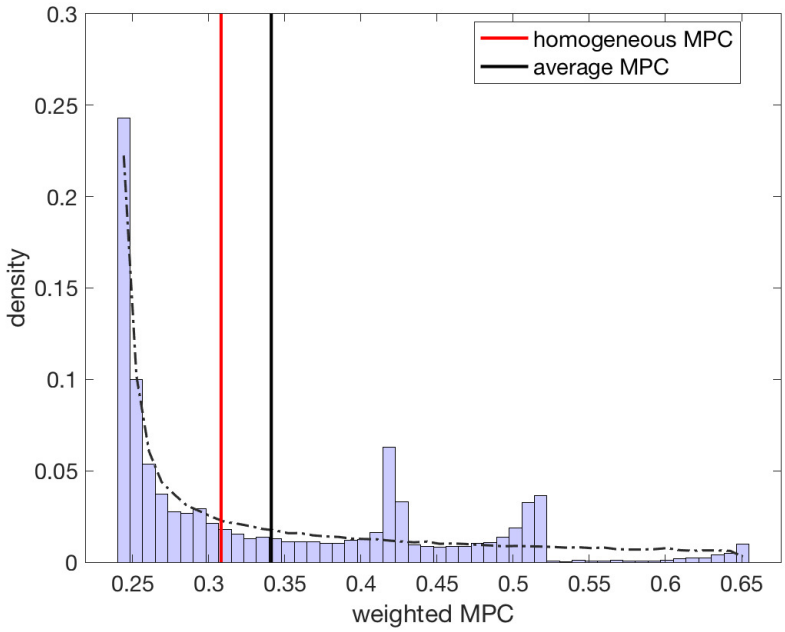
5.1 The Distribution of Marginal Propensities to Consume

As for virtually all the empirical variables considered in this paper, we define total expenditures as in Parker et al. (2013). Motivated by the simulation studies shown in Section 3.2 and the literature (see footnote 13), we set the fuzziness parameter m to 1.8. Following Kaplan and Violante (2014), who show that properly accounting for outliers reduces the homogeneous rebate coefficient, while increasing precision, we drop the top and bottom 1.5% of consumption changes.[22] After solving the algorithm for G ranging from 2 to 10, the gap statistic suggests that 5 is the optimal number of groups. For each household that receives the rebate, we compute the weighted average MPC, using the household-specific weights and the group-specific MPCs estimated by the algorithm.[23] Figure 2 shows the distribution of this object for those that received the rebate.

[22] This is the only way in which our sample departs from Parker et al. (2013), and explains why the homogeneous MPC we estimate for total consumption differs from theirs.

[23] Appendix Section D shows the distribution of the modal MPC – the one associated with the single highest weight at the individual level. The majority remains at an MPC of .25.

Figure 2: Estimated distribution of MPCs out of the tax rebate



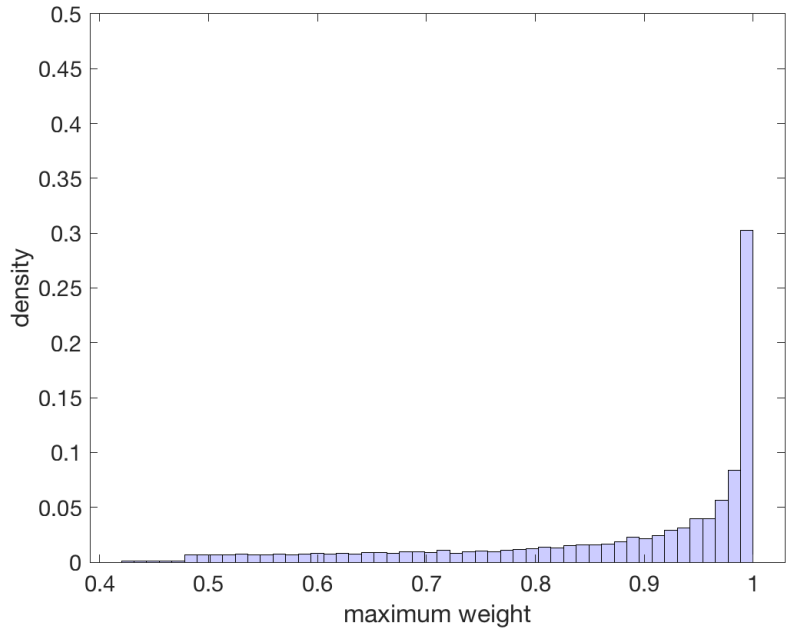
Notes: Figure 2 plots a histogram (light blue bars) of the estimated distribution of MPCs among households that received the rebate for total expenditures, defined as in Parker et al. (2013). The homogeneous MPC (vertical red line) is estimated assuming a homogeneous response to the tax rebate, also as in Parker et al. (2013) and following Equation 13. For each household we compute the weighted MPC, weighted across groups $g \in G$. The black vertical line shows the average weighted MPC in our sample. The dash-dotted line overlays data simulated from a Beta distribution, shifted to lie on the closed interval $[.245, .651]$, fitting our MPC distribution, with parameters 0.326 and 1.036.

The vast majority of households display a relatively low (but certainly non-negligible) MPC (~ 0.25), and the share of households with higher MPCs slowly decays as the MPC increases. While under this specification no household can be strictly defined as hand-to-mouth (MPC = 1), the majority of the sample exhibits a sizable propensity to consume; our findings suggest that most households consume at least part of the rebate.

We also document how, aggregating the individual-level responses, we obtain a larger propensity to consume than when running the homogeneous regression, as shown by the black and red vertical lines respectively. In Appendix D.4 we provide intuition for this result. The discrepancy is not driven by group-specific variation in the rebate, but by the properties of the joint distribution of the rebate, the controls, and the estimated MPC distribution. Our flexible approach allows us to account for non-linear heterogeneous relationships, which matter for aggregated responses.[24]

[24]In a model with heterogeneous effects, it is not generally true that an estimated homogeneous effect is equal to the weighted average of the heterogeneous effects. The sign, and the size, of this discrepancy, however, is a complicated object which depends on the joint correlation between treatment effects and covariates. In unreported results we find that individual controls do not separately account for the discrepancy. We check this by shutting down the contribution of each control in Equation (24).

Figure 3: Estimated distribution of maximum household weights



Notes: Figure 3 plots a histogram of modal estimated household weights for total expenditures, defined as in Parker et al. (2013).

Figure 3 depicts the distribution of estimated individual modal weights. Some households' assignments are estimated with near certainty (those with a maximum weight of ~1). The distribution of weights, however, is clearly quite different from the binary assignment that would result from HKM. Our weighted approach therefore allows us to recover the smoother distribution shown in Figure 2.

In Table 4 we show whether the estimated MPCs are statistically different from one another. In Table 4a we make use of the analytical formulas outlined in Theorem 6 to compute Wald tests of pairwise equality across MPCs. Groups 2 and 5 (ordered from lowest to highest MPC), whose MPCs are not statistically significantly different from zero, are also those with the smallest share of households.[25] The FCM standard errors are larger than standard errors on the equivalent weighted least squares regression, in which weights are taken as given, exactly because they take into account that the weights are estimated endogenously. Table 4b shows that - when group assignment is taken as given -all MPCs, except one, are statistically different from zero in this framework. Moreover, various MPC groups are statistically different from each other, at least at the 68% confidence level. Appendix Section D further shows that the *distribution* of MPCs is largely invariant when re-estimated on bootstrap samples drawn from the data.[26]

[25] 4% and 3% of rebaters assign maximum weight on groups 2 and 5, respectively.

[26] In particular, we repeat the estimation of the distribution of MPCs out of total expenditures, with 5

Table 4: Test for MPC equality

(a) Analytical standard errors

	MPC				
	0.24	0.29	0.42	0.52	0.65
0.24	6.66				
	(0.01)				
0.29	0.01	0.23			
	(0.92)	(0.63)			
0.42	1.04	0.05	3.33		
	(0.31)	(0.82)	(0.07)		
0.52	3.79	0.18	0.17	7.60	
	(0.05)	(0.67)	(0.68)	(0.00)	
0.65	0.35	0.16	0.14	0.04	0.82
	(0.55)	(0.69)	(0.71)	(0.82)	(0.37)

(b) Conditional on FCM weights

	MPC				
	0.24	0.29	0.42	0.52	0.65
0.24	84.5				
	(0.00)				
0.29	0.03	1.20			
	(0.87)	(0.27)			
0.42	8.49	0.24	51.8		
	(0.00)	(0.62)	(0.00)		
0.52	23.3	0.71	1.45	92.1	
	(0.00)	(0.40)	(0.23)	(0.00)	
0.65	1.79	0.81	0.55	0.19	4.61
	(0.18)	(0.37)	(0.46)	(0.66)	(0.03)

Notes: Total expenditures. The two tables show F-statistics from pairwise two-sided Wald tests of equality across MPCs (the diagonals shows tests of equality with zero). Table 4auses the standard errors outlined in Theorem 6. Table 4b repeats the exercise, taking the weights as given. These are equivalent to weighted least squares estimates where the weights are taken as given by those in Equation (4), raised to the power m as described in Theorem (2). Therefore, to run the tests in Table 4b, we replicate the sample by the number of groups and estimate $\Delta C_j = \beta' W_j + \sum_{g \in G} \left(\theta_g \mathbf{1}\left[j \in g\right] R_j + \alpha_g \mathbf{1}\left[j \in g\right] \right) + \epsilon_j$ via weighted least squares, with standard errors corrected for heteroskedasticity, and compute the Wald tests. P-values are reported in parentheses.

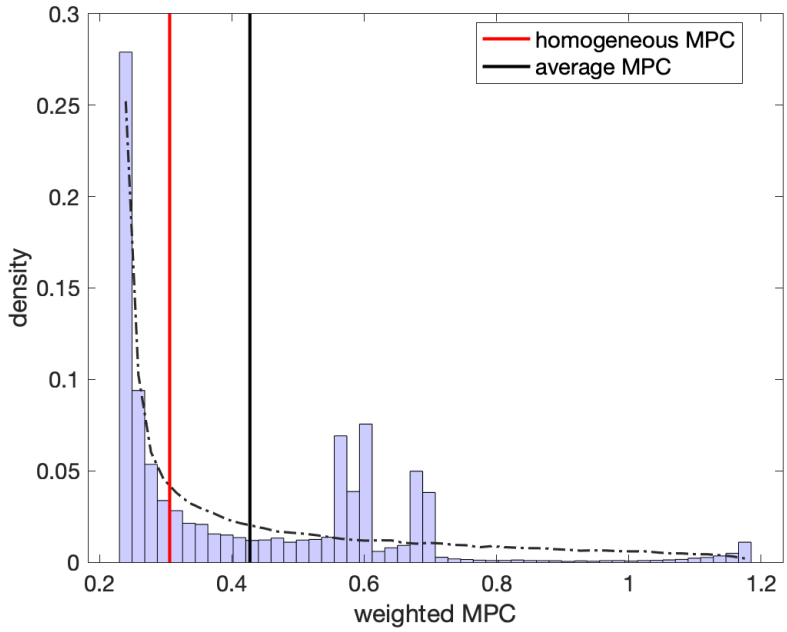
The flexibility of the FCM methodology allows us to nest instrumental variable estimation. This is particularly relevant in our framework, since the exogenous source of the transitory income shock is driven by the random timing of the rebate receipt, but the value of the rebate itself may be endogenous. We therefore follow the literature and instrument the tax rebate with an indicator function for its receipt. In this "TSLS" specification, we first regress the rebate value on a rebate indicator and the same controls as in Equation 14, and then use the predicted value in the second stage. Figure 4 plots the resulting distribution of weighted MPCs, and shows how it remains qualitatively unchanged relative to the OLS specification. If anything, instrumentation uncovers a small portion of households that consume the rebate in its entirety and that even display an MPC slightly larger than 1.[27] Moreover, the gap between aggregated and homogeneous response is even larger than in OLS.

Another concern raised by Kaplan and Violante (2014) is the interpretability of the re-

groups, over 100 samples obtained with bootstrap with replacement. We find that the average quantiles across bootstraps are very close to those estimated in the baseline sample, and fairly stable across bootstraps.

[27] In this setting, understanding the direction of bias in the OLS specification is difficult, since group membership is estimated simultaneously with the parameters. However, the correlation of individuals' estimated weighted MPCs across specifications is .91, and the rank correlation is .94, suggesting little movement in group membership by MPC.

Figure 4: Estimated distribution of MPCs out of the tax rebate: two-stage least squares



Notes: Figure 4 plots histogram (light blue bars) of the estimated distribution of MPCs for total expenditures, defined as in Parker et al. (2013), using the two-staged least squares specification. The homogeneous MPC (vertical red line) is estimated assuming a homogeneous response to the tax rebate, also as in Parker et al. (2013) and following Equation (13). We estimate the models with $G = 5$, to allow direct comparability with the distribution of MPCs out of total expenditures. For each household we compute the weighted MPC, weighted across groups $g \in G$. The black vertical line shows the average weighted MPC in our sample. The dash-dotted line overlays a histogram of data generated from a Beta distribution, shifted to lie in the closed interval $[.240, 1.18]$, fitting our MPC distribution, with parameters 0.322 and 1.203.

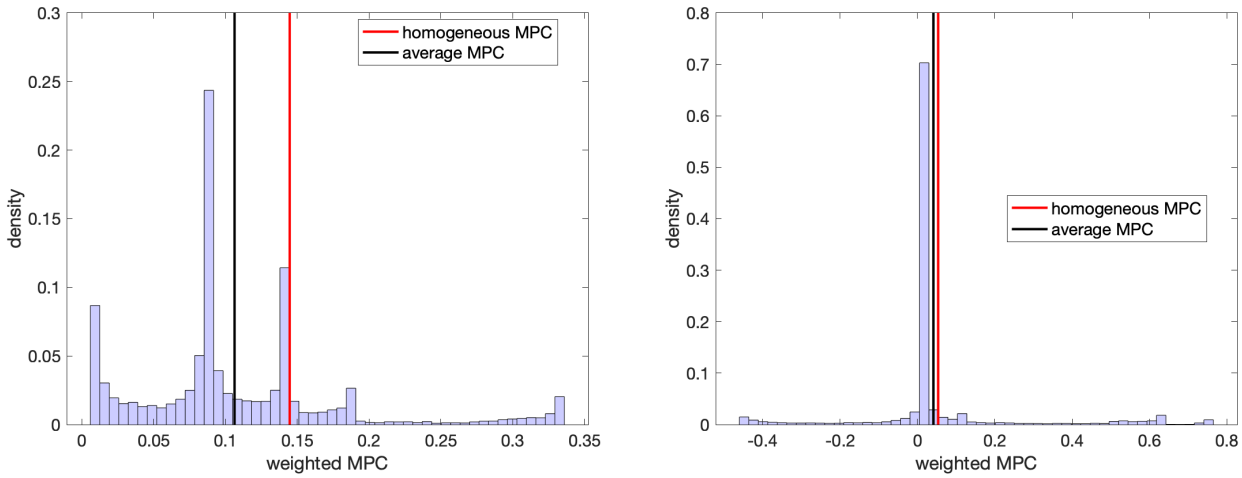
bate coefficient as a marginal propensity to consume. The concern arises from the observation that some households in the control group never receive a rebate (possibly because they have different characteristics, like higher income), some households in the control group have already received the rebate, and some households might anticipate receiving the rebate in the future. In Appendix Section D.3, we show that the estimated distribution maintains its main properties when we (i) drop households that never get the rebate and (ii) include lagged values of rebate.

5.2 The MPC Distribution for Different Consumption Goods

We have shown how households differ with respect to their propensity to consume the rebate. How does the distribution of these propensities change across consumption goods? The granularity of the CEX data allows us to tackle this question, while our approach allows us to explore how good-specific MPCs vary at the household level.

First, in the left panel of Figure 5, we show the weighted MPC distribution out of

Figure 5: MPCs out of the tax rebate: nondurables (left) and durables (right)



(a) Nondurables (b) Durables

Notes: Nondurable goods are defined, following Parker et al. (2013), as strictly nondurables (Lusardi (1996)) plus apparel goods and services, health care expenditures (excluding payments by employers or insurers), and reading material (excluding education). The homogeneous MPC (red line) is estimated assuming homogeneous response to the tax rebate. For each household we compute the weighted MPC, weighted across groups $g \in G$. The black line shows the average weighted MPC in our sample. We estimate the models with $G = 5$, to allow direct comparability with the distribution of MPCs out of total expenditures. We follow Coibion et al. (2017) and define durables as durable health expenditures, entertainment durables, furniture, jewelry, durable personal care, vehicle purchases, durable vehicle expenditures, housing durable expenditures (e.g., maintenance and repair commodities such as paint, materials.).

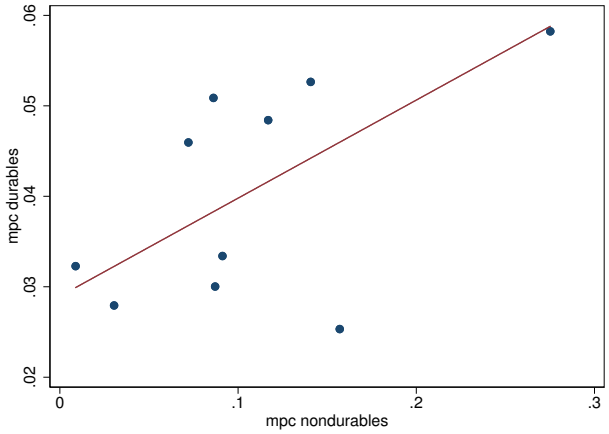
nondurable goods.[28] As expected, the distribution is shifted to the left with respect to the distribution corresponding to total expenditures in Figure 2, as nondurable goods account for, on average, only 57% of household total expenditures.

The vast majority of households consume a value of nondurables consistent with the annuity value of the rebate, as suggested by the Permanent Income Hypothesis (Friedman (1957)): between 71% and 92% of households have an MPC that is not statistically distinguishable from zero. A non-negligible portion of households, however, continue to display relatively large propensities to consume nondurable goods. The heterogeneity in nondurable MPCs is not only economically meaningful, but also statistically significant. In Appendix (D.1) we show that nearly all the estimated MPCs are statistically different from each other. Instrumenting the rebate with the rebate receipt indicator slightly increases both the mass and the values at the right tail, similar to the results for total expenditures.

As shown in the right panel of Figure 5, 86% of households are estimated not to change their durable expenditures in response to the rebate; their weighted MPC is below 0.05

[28]Nondurable goods are defined, following Parker et al. (2013), as strictly nondurables (Lusardi (1996)) plus apparel goods and services, health care expenditures (excluding payments by employers or insurers), and reading material (excluding education).

Figure 6: The correlation of MPCs across consumption goods



Notes: Figure 6 shows a binscatter of household MPC estimates for durables against nondurables. Each dot shows the average weighted MPC out of durable goods for each decile of the distribution of weighted MPCs out of nondurable goods.

and the associated group-specific MPC is statistically indistinguishable from the annuity value of the rebate. Moreover, 92% of the households have a modal MPC that is not statistically distinguishable from 0.05. The remaining MPCs, however, are 0.64 and 0.76.[29] The dichotomy of this MPC distribution stems directly from the specific features of durable goods. The discreteness of large purchases implies lumpy adjustment, and is consistent with the fact that most households either use most of the rebate to purchase durables, or do not adjust at all.

Finally, we check directly if households with high propensities to consume nondurable goods are also more likely to consume durable goods after receiving the rebate. The findings shown in Figure 6 suggest that this is the case. While we can rule out substitution between goods, the estimated complementarity - at the margin - is, however, quantitatively small. The correlation between households-level weighted MPCs out of nondurable goods with those for durables is 0.04, significant at the 5% level, while the rank correlation is 0.03. Albeit small, the complementarity might signal the presence of heterogeneous preferences or a small share of "spender" types, who are more prone to adjust any type of consumption in response to transitory income shocks. While the structure of our data does not allow us to draw conclusions regarding permanent unobserved heterogeneity in MPCs, we can investigate what observable characteristics explain the estimated MPC distributions that we recover. We tackle this issue in the next section.

[29] In line with the tendency shown for different consumption categories, estimating durable MPC with TSLS uncovers a group with larger propensity, up to 1.20.

5.3 What Drives MPC Heterogeneity?

Our approach uncovers the distribution of marginal propensities to consume without needing to take a stance on its observable drivers. Nevertheless, we can use the estimated distribution to understand how MPCs correlate with observable characteristics. This approach would be nearly impossible using existing approaches, since estimating MPCs for different observable subgroups (e.g., cut by age, income, wealth, and other observables simultaneously) would come at the cost of substantial loss of statistical power.

While many observables individually correlate with the MPCs, only three of them remain statistically significant explanatory variables even after the inclusion of additional drivers. We focus on them in this section and report individual correlations in Appendix Section D.[30]

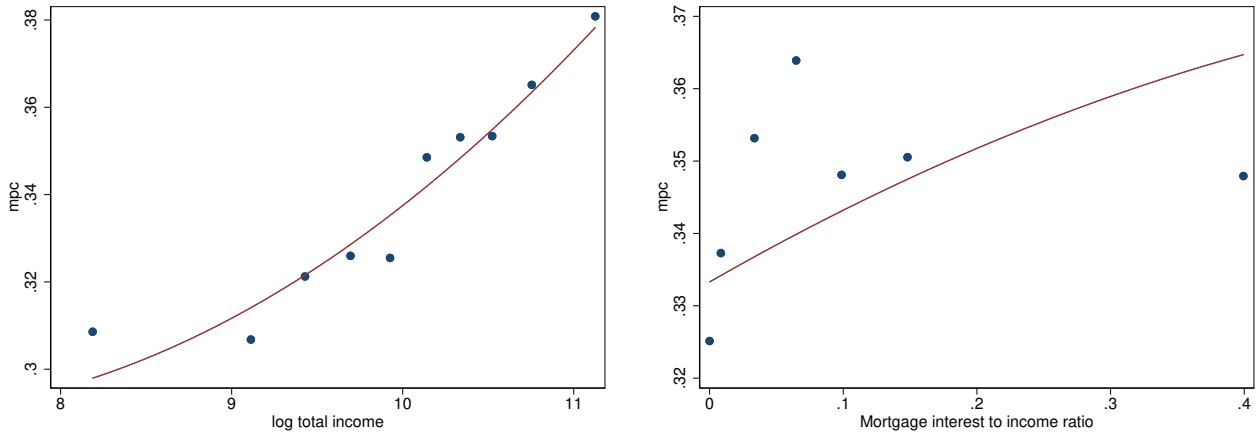
First, we find that high-income households have a greater marginal propensity to consume. While this is true for total income — defined as the sum of salary, financial and business income – it does not hold for salary income, once we control for the former. This effect does not seem to be driven by a particular category of households, such as entrepreneurs or investors (for example, those with a positive business or financial income), but rather by the intensive margin of total income. We find that a 1% increase in total income is associated with and increase in the MPC by 2 cents for each dollar of rebate; put differently, a 5 percent increase in income predicts 1 standard deviation increase in the MPC. While some studies find that low-income households have a higher marginal propensity to spend,[31] others are in line with our findings.[32] It should also be noted that income is measured in the CEX over the past 12 months and only in the first interview, thus making it less suited to measure transitory income fluctuations. The positive correlation between income and MPCs does not only hold for total expenditures, as shown in the left panel of Figure 7, but also for nondurable MPCs, even when controlling for additional covariates. In contrast, it holds only mildly and unconditionally for durable

[30] In unreported results we also show that our findings are virtually unchanged when considering the weighted MPC distribution estimated via TSLS, or the modal MPC distribution. Correlations with observables are also robust to the exclusion of observations associated with statistically insignificant MPCs out of total expenditures.

[31] For instance, Johnson et al. (2006) for the 2001 tax rebate and Jappelli and Pistaferri (2014), with respect to cash on hand, for Italian data on reported MPCs.

[32] Kueng (2018) studies consumption responses to regular and predetermined payments from the Alaska Permanent Fund and finds that MPCs monotonically increase with income. Misra and Surico (2014) also find that median income is higher at the top of the conditional distribution of consumption changes, which they find to be associated with higher propensities to consume, although the overall relationship is U-shaped. Shapiro and Slemrod (2009) use data on self-reported propensity to spend the 2008 rebate to show that low-income individuals were more likely to pay off debt. They also find that 21% of households making more than $75,000 of total annual income reported to mostly spend the rebate, compared to 18% for households with total income below $20,000.

Figure 7: Correlation of MPCs with total income and mortgage interest



(a) Total income

(b) Mortgage interest

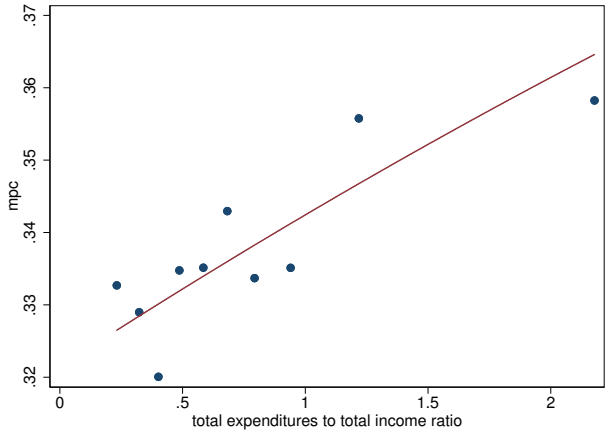
Notes: Figure 7shows the binscatter of MPCs agains total income (left panel) and mortgage to interest income ratios (right panel). Each dot shows the average weighted MPC out of total expenditures for each decile of the distribution of lagged log total income (left panel) and for each decile of the distribution of the ratio between mortgage interest payments and total income (right panel). The red line shows the quadratic fit. Log of total income takes a value of 0 when total income is 0 or negative. The mortgage interest to income ratio is winsorized at the top 99%.

MPCs.

The second important dimension is whether the household owns a home, and if so, whether the household has a mortgage. Homeowners are found to have greater MPCs, a result that echoes Parker et al. (2013). Furthermore, we find that having a mortgage is associated with an even higher propensity to consume, after we control for other drivers. The intensive margin of mortgage value also seems important. We examine the ratio between mortgage interest payments and total income, and find that a one percentage point increase in this ratio predicts 9 additional cents to be spent in total expenditures for each rebate dollar. The right panel of Figure 7 visually combines the extensive and intensive margin of mortgage status.

The last relevant observable is the average propensity to consume (APC). Empirically, we define the APC as the ratio between lagged consumption and lagged total income. As previously mentioned, we consider income as measured in the first interview for each households, and it refers to the previous 12 months. We lag expenditures to avoid a mechanical positive correlation with the MPC. To ensure stability of APCs, we average expenditures over all the available lagged quarters at the household-level, but the results are virtually unchanged if we only consider the first lag. Households that spent 1 percentage point more of their income before receiving the rebate spent 4 additional cents out of each rebate dollar. This effect is significant also for nondurable MPCs and conditional on

Figure 8: Marginal and average propensities to consume



Notes: Figure 8 shows a binscatter of the estimated MPCs against the household APCs, measured as mean lagged consumption relative to lagged total income. Each dot shows the average weighted MPC out of total expenditures for each decile of the distribution of the ratio between total expenditures and total income. The red line shows the quadratic fit. Expenditure rates are winsorized at 300%, corresponding to roughly the 97th top percentile. Observations with negative expenditure rates are dropped (0.03% of the sample).

a wide array of controls.[33] Figure 8 shows how this relationship is effectively linear.

We regard this result as particularly useful for disciplining macro models of household consumption. First, the average propensity to consume can be easily computed in a large number of micro datasets with minimal information. Second, this correlation can be directly tested in even the simplest of consumption/savings models. Yet, different models will have strikingly different implications for this moment. Consider the workhorse life-cycle model with incomplete markets. Households are born with zero assets and cannot borrow. Early on in the life cycle they are hand-to-mouth (APC $= 1$) and they display a large MPC. As they move up the income ladder, they start saving in order to accumulate a buffer stock. The APC starts falling, and so does the MPC. This behavior generates a positive correlation between the APC and MPC across the working-age population. How the remaining part of the population is modeled is important for the correlation. If agents are infinitely lived, they will save until a certain target wealth and then stop saving. As they approach the target, the MPC gradually falls towards the annuity value of the transitory income shock, while the APC converges to 1. In the population, this implies an ambiguous correlation between APC and MPC, in contrast with our results. In a life cycle model, instead, households start dissaving as they approach their death. This implies they display an APC > 1. Similarly, they are more responsive to transitory income shocks, given

[33] A 1 percentage point increase in total expenditures' APC predicts 2 additional cents per rebate dollar were spent in nondurables. This effect goes up to when considering the ratio between nondurable expenditures and total income.

the increasingly lower effective discount factor. This model has therefore the potential to generate a positive correlation between MPC and APC across the entire population.[34]

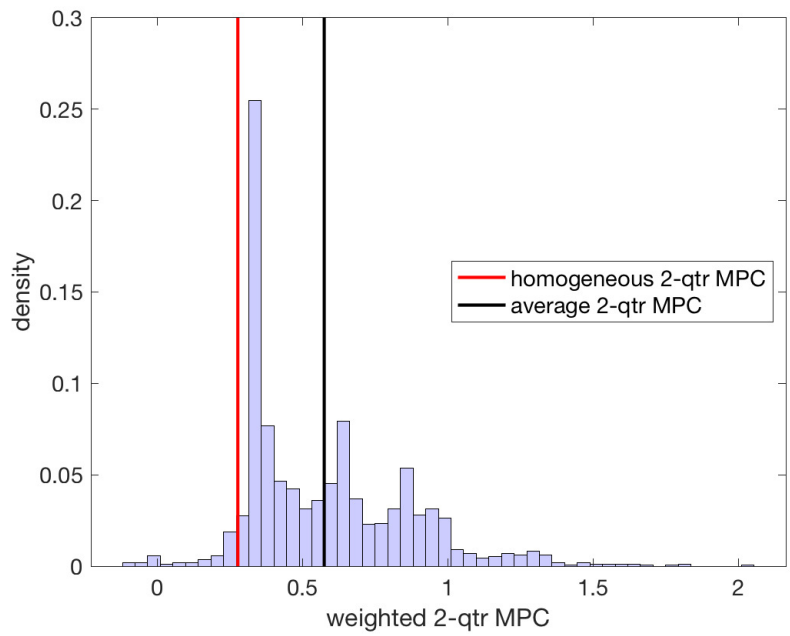
Finally, it is worth mentioning that all the observable drivers mentioned in this section — as well as other household characteristics that do not strongly correlate with the MPC — explain a relatively small portion of the variance of the weighted MPC distribution. Indeed, our best linear regression framework of weighted MPC on observable characteristics delivers an R^2 of 13%. This could be partly explained by non-linear relationships that are difficult to parametrize. Moreover, the CEX contains only sparsely populated information on wealth. In the Appendix, we show the relationship between the MPC and liquid wealth, aware of the potential nonresponse bias highlighted by Parker et al. (2013). We refrain from showing any relationship with total wealth, given the lack of reliable data. While these unobservable — within our dataset — characteristics could potentially explain some of the variation in MPCs, our results nevertheless suggest the presence of unobserved or latent drivers in MPC heterogeneity, especially since some of those latent characteristics may drive the observables we analyze in the first place. Our approach is able to uncover the full MPC distribution, including its latent part, but we are not able to clearly identify its source in the form of, for instance, preference heterogeneity. Some survey datasets try to directly to uncover these features. Parker (2017) finds that the majority of consumption responsiveness to the tax rebate, in the Nielsen data, is driven by a measure of impatience, defined as households reporting to be "the sort of people who would rather spend money and enjoy it today or save more for the future". Alternatively, a long panel data structure could allow one to draw conclusions on the permanent component of the MPC heterogeneity, as well as the evolution of the MPC distribution over the business cycle. The application of our framework to these questions is left to future research.

6 Aggregate partial equilibrium effects of the 2008 ESA

In this Section, we estimate the partial equilibrium (PE), aggregate response to the 2008 tax rebate based on our estimated heterogeneous coefficients. For this exercise, we use a lagged specification which takes into account the possible persistent effects of the rebate receipt, as in Parker et al. (2013). In particular, we estimate the following model:

[34]Retired households, on the other hand, are wealthier, which negatively affects the MPC in this class of models. Eventually, the correlation between MPC and APC depends on the quantitative properties of the model.

Figure 9: Estimated distribution of total 2-quarter effect of the tax rebate



Notes: Figure 9 plots histogram (light blue bars) of the estimated distribution of the total effect of the 2008 ESA for total expenditures, defined as in Parker et al. (2013), using the lagged specification in Equation 15. The homogeneous MPC (vertical red line) is estimated assuming a homogeneous contemporaneous and homogeneous lagged response to the tax rebate, also as in Parker et al. (2013). For each household we compute the weighted MPC, weighted across groups $g \in G$. The black vertical line shows the average weighted MPC in our sample.

$$\Delta C_j = \beta' W_j + \sum_{g \in G} \left(\theta_g \mathbf{1}\left[j \in g\right] R_j + \theta_g^{\text{lag}} \mathbf{1}\left[j \in g\right] R_j^{\text{lag}} + \alpha_g \mathbf{1}\left[j \in g\right] \right) + \epsilon_j \tag{15}$$

where the coefficient θ_g^{lag} represents the lagged effect of the rebate for group g.[35] We do not force a household to remain in a particular group in each period. To correctly estimate the cumulative response to the rebate, we therefore track individual weights over the two quarters following the rebate. We use these to construct the individual 2-quarter total effect of the rebate, by adding twice the weighted contemporaneous rebate coefficient to the weighted lagged coefficient.[36]

Figure 9 plots a histogram of this object among those who received the rebate. Relative to the baseline results depicted in Figure 2, the distribution spreads out, with some households having a total effect near zero. Moreover, as depicted in Figure 9, the es-

[35]See the Appendix, Section D.3 for further discussion of this specification.

[36]For example, a household may be categorized to be in some group a in the period in which they receive the rebate, and then in some group b the period after they receive the rebate. For such an individual, we construct the individual 2-quarter total effect of the rebate by adding twice the contemporaneous rebate coefficient for group a to the lagged rebate coefficient of group b.

timated partial equilibrium effect of the tax rebate doubles relative to its homogeneous counterpart, from .28 to .58.

7 Conclusion

We develop a flexible approach to uncover latent heterogeneity in cross section and short panel data, and use it to estimate heterogeneity in the marginal propensity to consume. We adapt the fuzzy C-means methodology, which jointly estimates group-specific coefficients and individual-specific membership weights, to a general regression framework. We motivate the use of fuzzy C-means by first demonstrating analytically that there always exists an m such that FCM is unbiased when $T = 1$ in a simple cluster means setting. We show equivalence between fuzzy C-means regression and the minimization of a nonlinear, weight-free, objective function, and establish asymptotic properties of the associated estimator using the fact that the new representation has a GMM formulation. In simulations, we show that the estimators preform very well, even when the data are not well-separated. As a further benefit, our estimator dramatically improves upon existing techniques in terms of computational speed. These features make fuzzy C-means regression well-suited to a wide range of economic problems featuring cross-section or short-panel data in the presence if unobserved heterogeneity.

We find that households display a considerable degree of heterogeneity in their marginal propensities to consume. Moreover, we show that different consumption goods are associated with different distributions, suggesting the need to take good-specific heterogeneity seriously in consumption/savings models. We do not find evidence of individual-level substitution across consumption goods in response to transitory income shocks, but rather a very mild positive correlation. Finally, we explore what observables best predict different portions of the MPC distribution. Our findings suggest that there is a tight relationship between marginal and average propensities to consume, which is easy to derive in many models of consumption behavior and yet has received relatively little attention. Since observable characteristics explain a minor portion of the estimated MPC heterogeneity, we posit that other latent factors might be important in determining marginal propensities to consume.

Finally, a few caveats are in order that highlight some open avenues for future work. Importantly, we measure the distribution of MPCs to the 2008 tax rebate. This means our estimated distribution uses a single cross-section of data during a recession; if an individual's MPC is a function of the aggregate state, extrapolating our estimates requires caution. Second, because our empirical setting is one in which individuals only experience

positive transitory shocks, we cannot speak to income windfalls, to which households may respond differently (Fuster et al. (2018)). However, the fuzzy C-means approach we develop can easily be extended to other datasets with suitably identified transitory income shocks, so that comparisons can be done. We leave such exercises open for future work.

References

AGUIAR, M., C. BOAR, AND M. BILS (2019): "Who Are the Hand-to-Mouth?" in *2019 Meeting Papers*, Society for Economic Dynamics, 525.

AUCLERT, A. (2019): "Monetary Policy and the Redistribution Channel," *American Economic Review*, 109, 2333–67.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

BEZDEK, J. (1981): *Pattern Recognition With Fuzzy Objective Function Algorithms*, Plenum Press.

BEZDEK, J. C., R. EHRLICH, AND W. FULL (1984): "FCM: The Fuzzy C-Means Clustering Algorithm," *Computers & Geosciences*, 10, 191 – 203.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2017): "Discretizing unobserved heterogeneity," *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

BONHOMME, S. AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184.

CARROLL, C., J. SLACALEK, K. TOKUOKA, AND M. N. WHITE (2017): "The Distribution of Wealth and the Marginal Propensity to Consume," *Quantitative Economics*, 8, 977–1020.

CHERNOZHUKOV, V., M. DEMIRER, E. DUFLO, AND I. FERNÁNDEZ-VAL (2017): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," *arXiv e-prints*, arXiv:1712.04802.

COIBION, O., Y. GORODNICHENKO, L. KUENG, AND J. SILVIA (2017): "Innocent Bystanders? Monetary policy and inequality," *Journal of Monetary Economics*, 88, 70–89.

CRAWLEY, E. AND A. KUCHLER (2018): "Consumption Heterogeneity: Micro Drivers and Macro Implications," *Danish National Bank Working Paper 129*.

DUNN, J. C. (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, 3, 32–57.

FAGERENG, A., M. B. HOLM, AND G. J. NATVIK (2016): "MPC Heterogeneity and Household Balance Sheets," Discussion Papers 852, Statistics Norway, Research Department.

FRIEDMAN, M. (1957): *A Theory of the Consumption Function*, Princeton University Press.

FUSTER, A., G. KAPLAN, AND B. ZAFAR (2018): "What Would You Do With $500? Spending Responses to Gains, Losses, News, and Loans," Staff Reports 843, Federal Reserve Bank of New York.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

HAYASHI, F. (2011): *Econometrics*, Princeton University Press.

JAPPELLI, T. AND L. PISTAFERRI (2014): "Fiscal Policy and MPC Heterogeneity," *American Economic Journal: Macroeconomics*, 6, 107–36.

JOHNSON, D. S., J. A. PARKER, AND N. S. SOULELES (2006): "Household Expenditure and the Income Tax Rebates of 2001," *American Economic Review*, 96, 1589–1610.

KAPLAN, G., B. MOLL, AND G. L. VIOLANTE (2018): "Monetary Policy According to HANK," *American Economic Review*, 108, 697–743.

KAPLAN, G. AND G. L. VIOLANTE (2014): "A Model of the Consumption Response to Fiscal Stimulus Payments," *Econometrica*, 82, 1199–1239.

KAPLAN, G., G. L. VIOLANTE, AND J. WEIDNER (2014): "The Wealthy Hand-to-Mouth," *Brookings Papers on Economic Activity*, 45, 77–153.

KUENG, L. (2018): "Excess sensitivity of high-income consumers," *The Quarterly Journal of Economics*, 133, 1693–1751.

LUSARDI, A. (1996): "Permanent Income, Current Income, and Consumption: Evidence from Two Panel Data Sets," *Journal of Business & Economic Statistics*, 14, 81–90.

MISRA, K. AND P. SURICO (2014): "Consumption, Income Changes, and Heterogeneity: Evidence from Two Fiscal Stimulus Programs," *American Economic Journal: Macroeconomics*, 6, 84–106.

NEWEY, W. K. AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," Elsevier, vol. 4 of *Handbook of Econometrics*, 2111 – 2245.

PAL, N. R. AND J. C. BEZDEK (1995): "On Cluster Validity for the Fuzzy C-Means Model," *IEEE Trans. Fuzzy Systems*, 3, 370–379.

PARKER, J. A. (2017): "Why Don't Households Smooth Consumption? Evidence from a 25*MillionExperiment*, *American Economic Journal: Macroeconomics*, 9, 153–83.

PARKER, J. A., N. S. SOULELES, D. S. JOHNSON, AND R. MCCLELLAND (2013): "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 103, 2530–53.

POLLARD, D. (1981): "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, 9, 135–140.

——— (1982): "A Central Limit Theorem for K-Means Clustering," *The Annals of Probability*, 10, 919–926.

SAHM, C. R., M. D. SHAPIRO, AND J. SLEMROD (2010): "Household Response to the 2008 Tax Rebate: Survey Evidence and Aggregate Implications," in *Tax Policy and the*

Economy, Volume 24, National Bureau of Economic Research, Inc, NBER Chapters, 69–110.

SHAPIRO, M. D. AND J. SLEMROD (2009): "Did the 2008 Tax Rebates Stimulate Spending?" *American Economic Review*, 99, 374–79.

SPIVAK, M. (1971): *Calculus On Manifolds: A Modern Approach To Classical Theorems Of Advanced Calculus*, Avalon Publishing.

TIBSHIRANI, R., G. WALTHER, AND T. HASTIE (2001): "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63, 411–423.

TORRA, V. (2015): "On the Selection of m for Fuzzy C-Means," .

WU, K.-L. (2012): "Analysis of Parameter Selections for Fuzzy C-Means," *Pattern Recognition*, 45, 407–415.

YANG, M.-S. (1994): "On Asymptotic Normality of a Class of Fuzzy C-Means Clustering Procedures," *International Journal of General Systems*, 22, 391–403.

YANG, M.-S. AND K. F. YU (1992): "On Existence and Strong Consistency of a Class of Fuzzy C-Means Clustering Procedures," *Cybernetics and Systems*, 23, 583–602.

YU, JIAN, CHENG, QIANSHENG, AND HUANG, HOUKUAN (2004): "Analysis of the Weighting Exponent in the FCM," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34, 634–639.

A Proofs

Proof of Proposition 1

Proof. Without loss of generality, we assume that $\zeta_1 = -\zeta_2$, so $(\zeta_1 + \zeta_2)/2 = 0$ (we can always demean all data before clustering). Denote the Gaussian with mean ζ_1 as G_1 and similarly G_2 for ζ_2, with $\zeta_1 < \zeta_2$, and denote their variance as σ^2. By symmetry, the groups are separated at 0, so values $y < 0$ are assigned to cluster 1 and $y > 0$ are assigned to cluster 2; $y = 0$ is a measure-zero event. Thus, the observations assigned to cluster 1 correspond to the portion of G_1 left of zero and the left tail of G_2. To compute ψ_1^*, it suffices to compute the mean over these two truncated normal distributions, weighted by

their relative contribution to the cluster's mass:

$$E\left[y \mid g\left(y\right)=1\right]=\frac{\Pr_{G_1}\left(g\left(y\right)=1\right)}{\Pr_{G_1}\left(g\left(y\right)=1\right)+\Pr_{G_2}\left(g\left(y\right)=1\right)}E_{G_1}\left[y \mid g\left(y\right)=1\right]$$
$$+\frac{\Pr_{G_2}\left(g\left(y\right)=1\right)}{\Pr_{G_1}\left(g\left(y\right)=1\right)+\Pr_{G_2}\left(g\left(y\right)=1\right)}E_{G_2}\left[y \mid g\left(y\right)=1\right],$$

where $g\left(y\right)$ denotes the group to which a value y is assigned. By symmetry, the total mass of the cluster is unity, so the relative contributions are simply $\Phi\left(-\zeta_1/\sigma\right)$ and $\Phi\left(-\zeta_2/\sigma\right)$ respectively, where Φ is the standard normal c.d.f. Finally, it remains to compute the means for each of the two truncated normals. Using standard results for the mean of the truncated normal distribution, with lower bound equal to $-\infty$ and upper bound equal to 0, we obtain the result,

$$\psi_1^*=\Phi\left(-\zeta_1/\sigma\right)\left(\zeta_1+\frac{-\phi\left(-\zeta_1/\sigma\right)}{\Phi\left(-\zeta_1/\sigma\right)}\right)+\Phi\left(-\zeta_2/\sigma\right)\left(\zeta_2+\frac{-\phi\left(-\zeta_2/\sigma\right)}{\Phi\left(-\zeta_2/\sigma\right)}\right),$$

with a symmetric argument giving a similar expression for ψ_2^*.

To conclude that ψ_1^* is negatively biased, note that in computing the mean over cluster 1, the right tail of G_1 (right of zero) has been replaced by an equal mass to the left of zero, shifting the overall mean to the left. ψ_1^* in general only recovers ζ_1^* by taking the limit as $\sigma \to 0$. □

Proof of Proposition 2

Proof. For any values of ρ_1 and ρ_2 and any $1 < m < \infty$, the maximal weight placed on a cluster for any value y is weakly lower under FCM than under HKM, in which it is always unity (with equality if and only if y is equal to either ρ_1 or ρ_2). This follows from the structure of the group weights,

$$\mu_g\left(y;\rho\right)=\left(\sum_{h=1}^{2}\frac{\left\|y-\rho_g\right\|^{2/(m-1)}}{\left\|y-\rho_h\right\|^{2/(m-1)}}\right)^{-1}=\left(1+\frac{\left\|y-\rho_g\right\|^{2/(m-1)}}{\left\|y-\rho_h\right\|^{2/(m-1)}}\right)^{-1}\leq 1, h\neq g.$$

Start by considering $\rho_1 = \psi_1^*, \rho_2 = \psi_2^*$. At these values (as at any others), the FCM weights on values corresponding to HKM cluster 1 are weakly lower than under HKM. Take any value $y_i \neq \psi_1^*$ assigned to cluster 1 (with weight 1) by HKM (so $y_i < 0$). Its weight on cluster 1 membership is now $1-\delta < 1$. There is a corresponding value, $-y_i$, assigned to cluster 2 by HKM (with zero weight on cluster 1), with weight on cluster 1 membership now given by $\delta > 0$, by symmetry. Jointly, the change in joint contribution to

cluster mean by these two points by changing HKM weights to FCM weights (evaluated at $\rho_1 = \psi_1^*, \rho_2 = \psi_2^*$) is $((1-\delta)-1)\, y_i + (\delta - 0)\,(-y_i) = -2\delta y_i > 0$, since $y_i < 0$. This argument can be repeated for all other values $y_j \neq \psi_1^*$ assigned to cluster 1 by HKM (for $y_j = \psi_1^*$ the weights are unchanged so the net effect is zero). The total change in cluster mean is given by

$$\int_{y<0, y\neq\psi_1^*} -2\delta\,(y_i)\, y_i P\,(dy) > 0,$$

where $\delta\,(y_i)$ expresses the change in weights from HKM to FCM as a function of y_i. Thus, the cluster mean computed based on FCM weights evaluated at $\rho_1 = \psi_1^*, \rho_2 = \psi_2^*$ is less than ψ_1^*. Of course, this is not the FCM optimum, since the weights were evaluated at different values (and thus the parameters do not constitute a fixed point). However, this argument can be repeated iteratively until a fixed point is obtained, since the weights become smoother still the closer ρ_1 becomes to ρ_2.

However, this argument began by evaluating the weights at the HKM parameters, so it remains to show that starting from values outwards of HKM does not yield a different conclusion. Consider first evaluating the weights at some arbitrary finite values $\rho_1' < \psi_1^*$ and $\rho_2' > \psi_2^*$. The maximal weights are still weakly lower relative to HKM, so after the first iteration, the FCM cluster means are again inside those of FCM, and the argument may continues exactly as above. Thus, ρ_1^* is located to the right of ψ_1^*, and ρ_2^* to the left of ψ_2^*.

For bounded m and finite σ^2, it is also the case that $\rho_1 \neq \rho_2 \neq 0$. This follows from Theorem 1 of Yang and Yu (1992), which shows that the objective function is lowered from the case of a single cluster mean by adding a second cluster mean. □

Proof of Theorem 1

Proof. We begin by showing that $\rho_i^*\,(m)$ is everywhere differentiable in m. ρ_i^* is implicitly defined by the moment equation

$$f_i\,(m,\rho) = \int_{-\infty}^{\infty} \left(1 + \frac{\left\| y_i - \rho_i^* \right\|^{\frac{2}{m-1}}}{\left\| y_i - \rho_j^* \right\|^{\frac{2}{m-1}}} \right)^{-m} (y - \rho_i^*)\, dP\,(y) = 0, i \neq j$$

(for a formal argument that the FCM clustering problem can be represented as a method of moments problem see (4)). By the implicit function theorem, since ρ_1^* and ρ_2^* are unique,

$$\frac{d\rho_i^*(m)}{dm} = \left[-\frac{\partial f}{\partial \rho'}(m, \rho^*(m))\right]^{-1}\left[-\frac{df}{dm}(m, \rho^*(m))\right],$$

where ρ stacks ρ_1, ρ_2 in a vector (and similarly f stacks f_1, f_2). Since ρ^* is assumed to be the unique solution to $f(m, \rho)$, the first term (the inverse of the Jacobian of the moments) exists. The second term can be simplified to

$$\int_{-\infty}^{\infty} 2(m-1)^{-2}\left[1 + \frac{\left\|y - \rho_i^* - \zeta_i\right\|^{\frac{2}{m-1}}}{\left\|y - \rho_j^* - \zeta_j\right\|^{\frac{2}{m-1}}}\right]^{-m} \ln\left[1 + \frac{\left\|y - \rho_i^* - \zeta_i\right\|^{\frac{2}{m-1}}}{\left\|y - \rho_j^* - \zeta_j\right\|^{\frac{2}{m-1}}}\right]$$
$$\times \frac{\left\|y - \rho_i^* - \zeta_i\right\|^{\frac{2}{m-1}}}{\left\|y - \rho_j^* - \zeta_j\right\|^{\frac{2}{m-1}}} \ln\left(\frac{\left\|y - \rho_i^* - \zeta_i\right\|}{\left\|y - \rho_j^* - \zeta_j\right\|^{\frac{2}{m-1}}}\right)(y - \rho_i^* - \zeta_i)\, dP(y).$$

For a given m (and thus ρ^*), the integrand is clearly finite for finite y except for at the point where $y = \rho_j^* + \zeta_j$ (and infinite y are probability zero since the variance of each Gaussian component is assumed to be finite). Further, denoting $x_j = \left\|y - \rho_j^* - \zeta_j\right\|^{\frac{2}{m-1}}$, $a_i = \left\|y - \rho_i^* - \zeta_i\right\|^{\frac{2}{m-1}}, \tilde{a}_i = \left\|y - \rho_i^* - \zeta_i\right\|$

$$\begin{aligned}
&\lim_{x_j \to 0}\left[1 + \frac{a_i}{x_j}\right]^{-m} \ln\left[1 + \frac{a_i}{x_j}\right]\frac{a_i}{x_j}\ln\left(\frac{\tilde{a}_i}{x_j}\right) \\
&= \lim_{x_j \to 0} \frac{\left[\frac{x_j + a_i}{x_j}\right]^{-m} a_i \ln\left[\frac{a_i + x_j}{x_j}\right]\ln\left(\frac{\tilde{a}_i}{x_j}\right)}{x_j} \\
&= \lim_{x_j \to 0} \frac{x_j^{m-1} a_i \left(\ln(a_i + x_j) - \ln(x_j)\right)\left(\ln(\tilde{a}_i) - \ln(x_j)\right)}{(x_j + a_i)^m} \\
&= \lim_{x_j \to 0} \frac{a_i}{(x_j + a_i)^m}\Big[x_j^{m-1}\left(\ln(a_i + x_j)\ln(\tilde{a}_i) - \ln(a_i + x_j)\ln(x_j)\right. . \\
&\qquad \left. - \ln(x_j)\ln(\tilde{a}_i) + \ln(x_j)\ln(x_j)\right)\Big]
\end{aligned}$$

The first part is clearly finite in the limit. We take the second part term-by-term. The first is clearly zero in the limit. The limit of the third term is zero by l'Hôpital's rule. The limits of the second and fourth terms are zero by double application of l'Hôpital's rule. Having

argued that the integrand is finite for all points with positive probability under P, the integral exists. Thus $\frac{\partial f}{\partial m}(m, \rho^*(m))$ exists everywhere, so $\frac{\partial \rho_i^*(m)}{\partial m}$ exists everywhere. Since $\rho_i^*(m)$ is a univariate function, existence of the derivative is sufficient for differentiability to hold. Since $\rho_i^*(m)$ is thus everywhere differentiable for $m \in (1, \infty)$, $\rho_i^*(m)$ is everywhere continuous. We know that $\lim_{m \to 1} \rho_1^*(m) = \psi_1^* < \zeta_1$, and that $\lim_{m \to \infty} \rho_1^*(m) = 0$ (Proposition 2). Note that given the normalization $(\zeta_1 + \zeta_2)/2 = 0$, $\zeta_1 < 0$. Therefore, by the intermediate value theorem, there exists some $\tilde{m} \in (1, \infty)$ such that $\rho_1^*(\tilde{m}) = \mu_1$. The same trivially holds for ρ_2^*. □

Proof of Theorem 2

Proof. The first point follows from simple algebra and the definition of μ_g^{reg}. In particular,

$$\begin{aligned} J_m^{reg}(\Pi, \mu^{reg}, \theta) &= \int \int \sum_{g=1}^{G} \mu_g^{reg,m}(y \mid x; \theta) \left\| y - \theta_g x \right\|^2 \Pi_{y|x}(dy \mid x) \Pi(dx) \\ &= \int \int \sum_{h=1}^{G} \left(\sum_{j=1}^{G} \frac{\|y - \theta_h x\|^{2/(m-1)}}{\left\| y - \theta_j x \right\|^{2/(m-1)}} \right) \left\| y - \theta_g x \right\|^2 \Pi_{y|x}(dy \mid x) \Pi(dx) \\ &= \int \int \left(\sum_{g=1}^{G} \left\| y - \theta_g x \right\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx), \end{aligned}$$

which is the formulation of $L_m^{reg}(\Pi, \theta)$. Parallel to the development of Yang and Yu (1992) for the cluster means case, we now show that minimization problem of J_m^{reg} is equivalent to that of L_m^{reg}, based on two lemmata from that paper.

Lemma 1. *(Yang and Yu (1992)) Let $p_g \geq 0$, $u_g > 0$ for $g = 1, \ldots, G$ such that $\sum_{g=1}^{G} p_g = 1$. Then*

$$\sum_{g=1}^{G} \left(\sum_{i=1}^{G} \frac{u_g^{1/(m-1)}}{u_i^{1/(m-1)}} \right)^{-1} u_i \leq \sum_{g=1}^{G} p_i^m u_i.$$

Proof. The proof is identical to that of Lemma 1 of Yang and Yu (1992). □

A simple modification of Lemma 2 of Yang and Yu (1992) completes the proof:

Lemma 2. *(Yang and Yu (1992)) Let θ^* be a minimizer of $L_m(\Pi, \theta)$ among all $\theta \in \Theta$. Then the pair $(\mu^{reg}(\theta^*), \theta^*)$ is a minimizer of $J_m^{reg}(\Pi, \mu, \theta)$ among all θ and weights μ.*

Proof. The proof follows directly from that of Lemma 2 of Yang and Yu (1992). □

Thus, the minimization of $L_m^{reg}(\Pi, \theta)$ is equivalent to the minimization of $J_m^{reg}(\Pi, \mu^{reg}, \theta)$, and we can restrict our attention to $L_m^{reg}(\Pi, \theta)$. □

Proof of Theorem 3

Proof. The proof is a straightforward extension of the proof of Theorem 1 in Yang and Yu (1992). Define $a(G) = \inf_\theta L_m^{reg}(\Pi, \theta)$ obtained for G groups. If Π is degenerate at some set of $T \times k$ matrices $\tilde{\theta}_g$, then $a(G) = 0$ and $\theta^* = \{\tilde{\theta}_1, \ldots \tilde{\theta}_G\}$. Therefore we can restrict our attention to non-degenerate Π. When $G = 1$, (9) reduces to $\int\int \|y - \theta x\|^2 \Pi(dy \mid x) \Upsilon(dx)$, which is the standard OLS objective function, which has the familiar solution $\theta_1^* = E[yx'] E[xx']^{-1}$ (the slightly different form accommodates y being $T \times 1$, T possibly greater than 1). Consider $G = 2$. Denote θ_{21} as the parameters for group 1 with $G = 2$, and let $\theta_{21} = \theta_1^*$, with θ_{22} arbitrary. Then

$$\begin{aligned} a(2) &\leq \int\int \left(\sum_{g=1}^{2} \|y - \theta_{2g} x\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\ &< \int\int \left(\sum_{g=1}^{1} \|y - \theta_g x\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\ &= a(1) < \infty, \end{aligned} \tag{16}$$

where the second inequality is strict since Π is not degenerate and $1 - m < 0$. Since $a(2) < \infty$, there exists $\theta^{(r)}(2) = \left(\theta_{21}^{(r)}, \theta_{22}^{(r)}\right)$ such that as $r \to \infty$,

$$\int\int \left(\sum_{g=1}^{2} \left\|y - \theta_{2g}^{(r)} x\right\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \to a(2).$$

We want to show that $\left\{\theta^{(r)}(2), r \geq 1\right\}$ is bounded. Suppose the statement is false, so there exists a subsequence $\theta_{21}^{(r_j)}$ such that $\left\|\theta_{21}^{(r_j)}\right\|$ goes to infinity. Then

$$
\begin{aligned}
a(2) &= \lim_{r_j \to \infty} \int \int \left(\sum_{g=1}^{2} \left\| y - \theta_{2g}^{(r_j)} x \right\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&\geq \liminf_{r_j \to \infty} \int \int \left(\sum_{g=1}^{2} \left\| y - \theta_{2g}^{(r_j)} x \right\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&\geq \int \int \liminf_{r_j \to \infty} \left(\sum_{g=1}^{2} \left\| y - \theta_{2g}^{(r_j)} x \right\|^{-2/(m-1)} \right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&\geq \int \int \liminf_{r_j \to \infty} \left\| y - \theta_{22}^{(r_j)} x \right\|^{2} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&\geq a(1),
\end{aligned}
$$

where the second inequality follows from Fatou's Lemma and the third uses the fact that $\left\|\theta_{21}^{(r_j)}\right\|$ goes to infinity. The result contradicts (16). Thus, $\left\{\theta^{(r)}(2), r \geq 1\right\}$ is bounded and there exist $\theta_{21}^*, \theta_{22}^*$ such that $\theta^{(r)}(2)$ converges to $\theta^*(2) = (\theta_{21}^*, \theta_{22}^*)$ along a subsequence, say r_j. Then for all $\delta > 0$, there exists r_0 such that for all $r_j > r_0$, $\left\|\theta^{(r_j)}(2) - \theta^*(2)\right\| \leq \delta$. Thus

$$
\begin{aligned}
&\left(\sum_{g=1}^{2} \left\| y - \theta_{2g}^{(r_j)} x \right\|^{-2/(m-1)} \right)^{1-m} \\
&\leq \max_{1 \leq i \leq 2} \left\| y - \theta_{2g}^{(r_j)} x \right\|^{2} \\
&\leq \left\| y - \theta_{21}^{(r_j)} x \right\|^{2} + \left\| y - \theta_{22}^{(r_j)} x \right\|^{2} \\
&\leq \left(\|y\| + \left\| \theta_{21}^{(r_j)} x \right\| \right)^{2} + \left(\|y\| + \left\| \theta_{22}^{(r_j)} x \right\| \right)^{2} \\
&\leq \left(\|y\| + \left\| \theta_{21}^{(r_j)} \right\| \|x\| \right)^{2} + \left(\|y\| + \left\| \theta_{22}^{(r_j)} \right\| \|x\| \right)^{2} \\
&\leq \left(\|y\| + \|\theta_{21}^*\| \|x\| + \|\delta \iota_{s \times k}\| \|x\| \right)^{2} + \left(\|y\| + \|\theta_{21}^*\| \|x\| + \|\delta \iota_{T \times k}\| \|x\| \right)^{2},
\end{aligned}
$$

where the third inequality follows from the triangle inequality, the fourth follows from Cauchy-Schwarz, and the last line follows from the triangle inequality and the fact that

$\delta \geq \left\|\theta^{(r_j)}(2) - \theta^*(2)\right\|$ implies $\left|\theta_{2g,tk}^{(r_j)}\right| \leq \left|\theta_{2g,tk}^*\right| + \delta$ for all g, t, k where t indexes dimensions of y and k indexes dimensions of x, which then implies $\left\|\theta_{2g}^{(r_j)}\right\| \leq \left\|\theta_{2g}^*\right\| + \|\delta \iota_{T\times k}\|$. By Assumption 2, the last line provides a bound in expectation for the left hand side. Finally, since the last line establishes a bounding function for $\left(\sum_{g=1}^{2}\left\|y - \theta_{2g}^{(r_j)}x\right\|^{-2/(m-1)}\right)^{1-m}$, the dominated convergence theorem shows that as r_j tends to infinity,

$$
\begin{aligned}
a(2) &= \lim_{r_j\to\infty}\int\int\left(\sum_{g=1}^{2}\left\|y - \theta_{2g}^{(r_j)}x\right\|^{-2/(m-1)}\right)^{1-m}\Pi_{y|x}(dy \mid x)\,\Pi(dx) \\
&= \int\int\left(\sum_{g=1}^{2}\left\|y - \theta_{2g}^{*}x\right\|^{-2/(m-1)}\right)^{1-m}\Pi_{y|x}(dy \mid x)\,\Pi(dx).
\end{aligned}
$$

This establishes that the infimum $a(2)$ is indeed obtained at $\theta^*(2)$, the limit of $\theta^{(r)}(2)$ for subsequence r_j (which exists). A similar argument can then be made sequentially for $G = 3, 4, \ldots$, so by mathematical induction, the theorem is therefore true for all $G = 1, 2, \ldots$. □

Proof of Theorem 4

Proof. We start by differentiating the integrand of $L_m^{reg}(\Pi, \theta)$ with respect to $\theta_{g,tk}$:

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_{g,tk}}\left(\sum_{h=1}^{G}\|y - \theta_h x\|^{-2/(m-1)}\right)^{1-m} \\
&= (1-m)\left(\sum_{h=1}^{G}\|y - \theta_h x\|^{-2/(m-1)}\right)^{-m}\frac{\partial}{\partial\theta_{g,tk}}\sum_{g=1}^{G}\|y - \theta_g x\|^{-2/(m-1)} \\
&= (1-m)\left(\sum_{h=1}^{G}\|y - \theta_h x\|^{-2/(m-1)}\right)^{-m}\frac{-2}{m-1}\|y - \theta_g x\|^{(1+m)/(1-m)}\frac{\partial}{\partial\theta_{g,tk}}\|y - \theta_g x\| \\
&= 2\left(\sum_{h=1}^{G}\|y - \theta_h x\|^{-2/(m-1)}\right)^{-m}\|y - \theta_g x\|^{-(1+m)/(m-1)}\frac{y_t - \theta_{g,(t)}x}{\|y - \theta_g x\|}(-x_k) \\
&= -2\left(\sum_{h=1}^{G}\|y - \theta_h x\|^{-2/(m-1)}\right)^{-m}\|y - \theta_g x\|^{-2m/(m-1)}\left(y_t - \theta_{g,(t)}x\right)x_k \\
&= -2\left(\sum_{h=1}^{G}\frac{\|y - \theta_g x\|^{2/(m-1)}}{\|y - \theta_h x\|^{2/(m-1)}}\right)^{-m}\left(y_t - \theta_{g,(t)}x\right)x_k,
\end{aligned}
$$

where $\theta_{g,(t)}$ denotes the row of θ_g corresponding to outcome y_t. Note that since these partial derivatives are continuous in θ (by inspection; see also Yang (1994) Lemma 2), the integrand is (continuously) differentiable in θ (Spivak (1971) Theorem 2.8). Moreover, $\left(\sum_{h=1}^{G} \|y - \theta_h x\|^{-2/(m-1)}\right)^{1-m}$ is Lebesgue-integrable for each θ as

$$\left(\sum_{h=1}^{G} \|y - \theta_h x\|^{-2/(m-1)}\right)^{1-m} \leq \sum_{h=1}^{G} \|y - \theta_h x\|^{-2(1-m)/(m-1)} = \sum_{h=1}^{G} \|y - \theta_h x\|^2$$

since $1 - m < 0$ and

$$\begin{aligned}
\sum_{h=1}^{G} \|y - \theta_h x\|^2 &\leq \sum_{h=1}^{G} (\|y\| + \|\theta_h x\|)^2 \\
&\leq \sum_{h=1}^{G} (\|y\| + \|\theta_h\| \|x\|)^2 \\
&= \sum_{h=1}^{G} \|y\|^2 + 2 \|\theta_h\| \|x\| \|y\| + \|\theta_h\|^2 \|x\|^2 , \qquad (17)
\end{aligned}$$

which is integrable by Assumptions 2.1 and 2.3. Moreover, (17) establishes a bounding function for the integrand in terms of θ. From these conditions, the dominated convergence theorem allows the interchange of differentiation and integration:

$$\begin{aligned}
\frac{\partial L_m^{reg}(\Pi, \theta)}{\partial \theta_{g,tk}} &= \frac{\partial}{\partial \theta_{g,tk}} \int \int \left(\sum_{g=1}^{G} \|y - \theta_g x\|^{-2/(m-1)}\right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&= \int \int \left(\frac{\partial}{\partial \theta_{g,tk}} \sum_{g=1}^{G} \|y - \theta_g x\|^{-2/(m-1)}\right)^{1-m} \Pi_{y|x}(dy \mid x) \Pi(dx) \\
&= E\left[\left(\frac{\partial}{\partial \theta_{g,tk}} \sum_{g=1}^{G} \|y_i - \theta_g x_i\|^{-2/(m-1)}\right)^{1-m}\right] \\
&= E\left[-2\left(\sum_{h=1}^{G} \frac{\|y_i - \theta_g x_i\|^{2/(m-1)}}{\|y_i - \theta_h x_i\|^{2/(m-1)}}\right)^{-m} \left(y_{it} - \theta_{g,(t)} x_i\right) x_{i,k}\right],
\end{aligned}$$

where we henceforth replace the Lebesgue integrals with expectations. Stacking the conditions vertically for row t of θ_g yields the $k \times 1$ vector

$$\frac{\partial L_m^{reg}}{\partial \theta'_{g,(t)}} = E\left[-2\left(\sum_{h=1}^{G} \frac{\|y-\theta_g x\|^{2/(m-1)}}{\|y-\theta_h x\|^{2/(m-1)}}\right)^{-m}\left(y_t - \theta_{g,(t)} x\right) x\right].$$

Proceeding likewise across $t = 1, \ldots, T$ and for $g = 1, \ldots, G$ yields $G \times T \times k$ conditions which θ^* must satisfy,

$$E\left[\left(\sum_{h=1}^{G} \frac{\|y_i-\theta_g x_i\|^{2/(m-1)}}{\|y_i-\theta_h x_i\|^{2/(m-1)}}\right)^{-m}\left(y_{it} - \theta_{g,(t)} x_i\right) x_i\right] = 0, \text{ for } g = 1, \ldots, G, t = 1, \ldots, T,$$

since θ^* minimizes $L_m^{reg}(\Pi, \theta)$. These $G \times T \times k$ equations constitute moment conditions for the $G \times T \times k$ free parameters in θ. Thus, the system of equations constitutes a just-identified GMM problem. □

Proof of Theorem 5

Proof. By Assumption 2.1, (y_i, x_i) are i.i.d. By Assumption 3, θ^* uniquely satisfies $\eta(\theta, y_i, x_i)$. As noted in the proof of Corollary 4, the moment conditions $\eta(\theta, y_i, x_i)$ are continuous for all $\theta \in \Theta$. Next, we show that the moments are bounded in expectation for all $\theta \in \Theta$ (the dominance condition). Observe that $\left(\sum_{h=1}^{G} \frac{\|y-\theta_g x\|^{2/(m-1)}}{\|y-\theta_h x\|^{2/(m-1)}}\right)^{-m}$ is bounded between zero and one (the supremum of the summation is infinity as the residuals $y - \theta_h x$, $h \neq g$ go to zero and the infimum is 1 as $y - \theta_h x$, $h \neq g$ go to infinity). So

$$\begin{aligned}
E\left[\sup_{\theta\in\Theta} \|\eta(\theta, y_i, x_i)\|\right] &\leq E\left[\sup_{\theta\in\Theta}\sup_{g} \left\|\left(y_i - \theta_g x_i\right) x_i'\right\|\right] \\
&= E\left[\sup_{\theta\in\Theta}\sup_{g} \left\|y_i x_i' - \theta_g x_i x_i'\right\|\right] \\
&\leq E\left[\sup_{\theta\in\Theta}\sup_{g} \left\|y_i x_i'\right\| + \left\|\theta_g x_i x_i'\right\|\right] \\
&\leq E\left[\sup_{\theta\in\Theta}\sup_{g} \|y_i\| \|x_i\| + \|\theta_g\| \|x_i\| \|x_i\|\right] \\
&< \infty,
\end{aligned}$$

where the third inequality follows from the triangle inequality, the fourth from Cauchy-Schwarz, and the final follows from Assumptions 2.1 and 2.3. These points jointly satisfy the requirements of standard GMM arguments, (e.g., Newey and McFadden (1994), p. 2121—2, Hayashi (2011) Proposition 7.7), so$\hat{\theta} \xrightarrow{p} \theta^*$. □

Proof of Theorem 6

Proof. First, we provide expressions for H to establish the continuous differentiability of $\eta(\theta, y_i, x_i)$ in θ. We focus on the cross-sectional case here ($T = 1$) for the sake of simplicity and in keeping with our empirical focus , but provide fully general expressions for panel data in Section B.1. Partition the blocks of H as

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1g} & \cdots & H_{1G} \\ \vdots & \ddots & & & \vdots \\ H_{g1} & & H_{gg} & & H_{gG} \\ \vdots & & & \ddots & \vdots \\ H_{G1} & \cdots & H_{Gg} & \cdots & H_{GG} \end{bmatrix},$$

where $H_{gh} = \frac{\partial^2 L_m^{reg}}{\partial\theta_g \partial\theta_h'}$,with $H_{gh} = H_{hg}'$ by symmetry of the Hessian. For the case where all coefficients are group-specific, it can be shown that

$$H_{gg} = E\left[x_i x_i' \left\{\frac{-2m}{m-1} A_i^{-m-1} \left(e_{i,g}\right)^2 C_{i,g}^2 + \frac{m+1}{m-1} A_i^{-m} C_{i,g}\right\}\right]$$
$$H_{gh} = E\left[x_i x_i' \left\{\frac{-2m}{m-1} A_i^{-m-1} C_{i,h} e_{i,h} e_{i,g} C_{i,g}\right\}\right], h \neq g,$$

where $e_{i,g} = y_i - \theta_g x_i$, $A_i = \Sigma_{g=1}^G \left\|e_{i,g}\right\|^{-2/(m-1)}$, $C_{ig} = \left\|e_{i,g}\right\|^{-2m/(m-1)}$. We also provide expressions for additional elements of the Hessian when there are covariates with common coefficients across groups, such that $\theta_{g,k} = \theta_{h,k} \equiv \theta_{\star,k}, h \neq g$. In this case,

$$\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,k}\partial\theta_{\star,k}} E\left[x_{i,k}^2 \left\{\frac{-2m}{m-1} A_i^{-m-1} B_i^2 + \frac{m+1}{m-1} A_i^{-m} \sum_{g=1}^G C_{i,g}\right\}\right]$$
$$\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,k}\partial\theta_{\star,l}} = E\left[x_{i,k} x_{i,l} \left\{\frac{-2m}{m-1} A_i^{-m-1} B_i^2 + \frac{m+1}{m-1} A_i^{-m} \sum_{g=1}^G C_{i,g}\right\}\right]$$
$$\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,k}\partial\theta_{g,l}} = E\left[x_{i,k} x_{i,l} \left\{\frac{-2m}{m-1} A_i^{-m-1} C_{i,g} e_{i,g} B_i + \frac{m+1}{m-1} A_i^{-m} C_{i,g}\right\}\right],$$

where $B_i = \sum_{g=1}^{G} \left[e_{i,g} C_{i,g} \right]$. By inspection, all elements of these Hessians are continuous in θ, since $e_{i,g}, A_i^{-m}, A_i^{-m-1}, C_{ig}, B_i$ are continuous in θ, and all elements of H are continuous functions of these objects.

Next, we establish the asymptotic normality of $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \eta\left(\theta, y_i, x_i\right)$. Since y_i, x_i are assumed to be jointly i.i.d., $\eta\left(\theta, y_i, x_i\right)$ is i.i.d. across observations, so by the Lindeberg-Levy central limit theorem,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \eta\left(\theta^*, y_i, x_i\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

where $V = E\left[\eta\left(\theta, y_i, x_i\right) \eta\left(\theta, y_i, x_i\right)'\right]$ is assumed to be positive definite in Assumption 4.4.

Combining these two results with the conditions of Assumption 4, the standard conditions for asymptotic normality of a GMM estimator are satisfied (e.g., Hayashi (2011) Proposition 7.10). Since the weighting matrix is the identity (the problem is just-identified),

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1} V H^{-1}\right).$$

□

B Supplemental theoretical results

In this section, we report extensions of our main theoretical results. First, we provide expressions for the Hessian to compute the asymptotic variance in the case of panel data. Second, we extend all regression results from the main text to a TSLS implementation. Third, we extend the fixed-T Hessians provided in the Appendix of Bonhomme and Manresa (2015) to the case of heterogeneous coefficients on regressors. Finally, we discuss the implications of including additional controls in the regression function for recovering the true heterogeneity of the data.

B.1 Hessian for panel data

The Hessian provided in the proof of Theorem 6 assumes $T = 1$, the cross-sectional data case that is the focus of our empirical study. However, the results of the paper hold in generality for $T > 1$. Here, we report the elements of the Hessian for the case of $T > 1$, corresponding to a panel structure. Without common coefficients, the second

partial derivatives are given by

$$
\begin{aligned}
\frac{\partial^2 L_m^{reg}}{\partial\theta_{g,kt}\partial\theta_{g,lt}} &= E\left[x_{i,k}\left\{-mA^{-m-1}\left(\frac{2}{m-1}\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,t}x_{i,l}\right)e_{i,g,t}C_{i,g}\right.\right.\\
&\quad\left.\left.-A^{-m}x_{i,l}C_{i,g}+A^{-m}e_{i,g,t}\left(\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,t}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A^{-m-1}C_{i,g}^2e_{i,g,t}^2-A^{-m}C_{i,g}+\frac{2m}{m-1}A^{-m}C_{i,g}^{\frac{2m-l}{m}}e_{i,g,t}^2\right\}\right]\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{g,kt}\partial\theta_{g,ls}} &= E\left[x_{i,k}\left\{-mA^{-m-1}\left(\frac{2}{m-1}\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,s}x_{i,l}\right)e_{i,g,t}C_{i,g}\right.\right.\\
&\quad\left.\left.+A^{-m}\times 0\times C_{i,g}+A^{-m}e_{i,g,t}\left(\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,s}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A^{-m-1}C_{i,g}^2e_{i,g,t}e_{i,g,s}+\frac{2m}{m-1}A^{-m}C_{i,g}^{\frac{2m-l}{m}}e_{i,g,t}e_{i,g,s}\right\}\right]\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{g,kt}\partial\theta_{h,lt}} &= E\left[x_{i,k}\left\{-mA^{-m-1}\left(\frac{2}{m-1}\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h,t}x_{i,l}\right)e_{i,g,t}C_{i,g}\right.\right.\\
&\quad\left.\left.+A^{-m}\times 0\times C_{i,g}+A^{-m}e_{i,g,t}\times 0\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A^{-m-1}C_{i,g}C_{i,h}e_{i,g,t}e_{i,h,t}\right\}\right], h\neq g\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{g,kt}\partial\theta_{h,ls}} &= E\left[x_{i,k}\left\{-mA^{-m-1}\left(\frac{2}{m-1}\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h,s}x_{i,l}\right)e_{i,g,t}C_{i,g}\right.\right.\\
&\quad\left.\left.+A^{-m}\times 0\times C_{i,g}+A^{-m}e_{i,g,t}\times 0\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A^{-m-1}C_{i,g}C_{i,h}e_{i,g,t}e_{i,h,s}\right\}\right], h\neq g.
\end{aligned}
$$

With common coefficients across groups, additional partial derivatives with respect to the common coefficients must be obtained. For this purpose, let $B_{it}=\sum_{g=1}^{G}\left[e_{i,g,t}C_{i,g}\right]$. Then the relevant derivatives are given by

$$
\begin{aligned}
\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,kt}\partial\theta_{\star,lt}} &= E\left[x_{i,k}\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{g=1}^{G}\left(\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,t}x_{i,l}\right)B_{it}\right.\right.\\
&\quad\left.\left.+A^{-m}\sum_{g=1}^{G}\left(-x_{i,l}C_{i,g}+e_{i,g,t}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,t}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}B_{it}^2+A_i^{-m}\sum_{g=1}^{G}\left(\frac{2m}{m-1}C_{i,g}^{\frac{2m-1}{m}}e_{i,g,t}^2-C_{i,g}\right)\right\}\right]\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,kt}\partial\theta_{\star,ls}} &= E\left[x_{i,k}\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{g=1}^{G}\left(\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,s}x_{i,l}\right)B_{it}\right.\right.\\
&\quad\left.\left.+A^{-m}\sum_{g=1}^{G}\left(0\times C_{i,g}+e_{i,g,t}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,s}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}B_{is}B_{it}+A_i^{-m}\sum_{g=1}^{G}\left(\frac{2m}{m-1}C_{i,g}^{\frac{2m-1}{m}}e_{i,g,s}e_{i,g,t}\right)\right\}\right]\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,k,}\partial\theta_{g,lt}} &= E\left[x_{i,k}\left\{\frac{-2m}{m-1}A^{-m-1}\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,t}x_{i,l}B_{it}\right.\right.\\
&\quad\left.\left.+A^{-m}\left(-x_{i,l}C_{i,g}+e_{i,g,t}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,t}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}C_{i,g}e_{i,g,t}B_{it}+A_i^{-m}\left(\frac{2m}{m-1}C_{i,g}^{\frac{2m-1}{m}}e_{i,g,t}^2-C_{i,g}\right)\right\}\right]\\
\frac{\partial^2 L_m^{reg}}{\partial\theta_{\star,kt}\partial\theta_{g,ls}} &= E\left[x_{i,k}\left\{\frac{-2m}{m-1}A^{-m-1}\left\|e_{i,g}\right\|^{-2m/(m-1)}e_{i,g,s}x_{i,l}B_{it}\right.\right.\\
&\quad\left.\left.+A^{-m}\left(0\times C_{i,g}+e_{i,g,t}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g,s}x_{i,l}\right)\right\}\right]\\
&= E\left[x_{i,k}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}C_{i,g}e_{i,g,s}B_{it}+A_i^{-m}\sum_{g=1}^{G}\left(\frac{2m}{m-1}C_{i,g}^{\frac{2m-1}{m}}e_{i,g,s}e_{i,g,t}\right)\right\}\right].
\end{aligned}
$$

B.2 Properties of TSLS FCM

In the text, we describe a TSLS-type procedure using FCM. Here, we describe its theoretical properties in detail. For the purposes of this development, we assume cross-sectional data, so $T = 1$ (y_i is a scalar). Let x^e denote the endogenous regressors of interest, and let

ω denote additional controls. Consider a homogeneous first-stage regression

$$x_i^e = \gamma z_i + \tau \omega_i + u_i, \tag{18}$$

for k_z instruments z with $k_z \geq k_e$, where k_e is the number of endogenous regressors x^e. We henceforth consider the just-identified case with a single endogenous regressor, $k_z = k_e = 1$, for economy of notation and in keeping with our empirical focus, but the results can be trivially extended to allow for additional dimensions, overidentification, and an arbitrary weight matrix W in the first-stage. We also assume that x^e has heterogeneous coefficients. Denote $\tilde{x}_i^e = \gamma z_i + \tau \omega_i$, the predicted first-stage values, and $\tilde{x}_i = \begin{pmatrix} \tilde{x}_i^e, & \omega' \end{pmatrix}'$, the vector of predicted endogenous regressors and exogenous controls (so $\tilde{x}^e$ is ordered first). Using these values, we define the FCM second-stage as

$$J_m^{TSLS} = E\left[\sum_{g=1}^{G} \mu_g^{TSLS,m}\left(y_i \mid \tilde{x}_i; \theta^{TSLS}\right) \left\| y_i - \theta_g^{TSLS} \tilde{x}_i \right\|^2\right],$$

where

$$\mu_g^{TSLS}\left(y_i \mid \tilde{x}_i; \theta^{TSLS}\right) = \left(\sum_{h=1}^{G} \frac{\left\| y_i - \theta_g^{TSLS} \tilde{x}_i \right\|^{2/(m-1)}}{\left\| y_i - \theta_h^{TSLS} \tilde{x}_i \right\|^{2/(m-1)}}\right)^{-1}, g = 1, \ldots, G.$$

Assumption 5. *1. $\left(y_i, x_i^e, w_i, z_i\right)$ are i.i.d. with probability measure $\tilde{\Pi}$ and $E\left[\varepsilon_i z_i\right] = 0$,*

2. The second moments of y and $\tilde{x}$ are finite under $\tilde{\Pi}$:

$$E\left[y_i^2\right] < \infty,\ E\left[\tilde{x}_i^2\right] < \infty,\ E\left[\tilde{x}_i y_i\right] < \infty,$$

3. Additionally, neither $\tilde{x}$ nor $\begin{pmatrix} z \\ \omega \end{pmatrix}$ is collinear,

$$rank\left(E\left[\tilde{x}_i \tilde{x}_i'\right]\right) = k,$$

$$rank\left(E\left[\begin{pmatrix} z_i \\ \omega_i \end{pmatrix}\begin{pmatrix} z_i \\ \omega_i \end{pmatrix}'\right]\right) = k - k_e + k_z.$$

Assumption 5 stipulates the TSLS assumptions and regularity conditions on the data. The relevance condition is incorporated in point 3.

Corollary 1. *Define* $L_m^{TSLS} = E\left[\left(\sum_{g=1}^{G}\left\|y-\theta_g^{TSLS}\tilde{x}\right\|^{-2/(m-1)}\right)^{1-m}\right]$. *If* γ *and* τ *are known and Assumption 5 holds, then*

1. L_m^{TSLS} *is equivalent to* J_m^{TSLS},
2. *There exists a solution to* L_m^{TSLS}, $\theta^{*,TSLS}$.

Proof. The results follow immediately from Theorems 2 and 3, simply replacing x with $\tilde{x}$. □

Corollary 1 establishes the existence of a solution to the FCM problem, $\theta^{*,TSLS}$, which minimizes L_m^{TSLS}. L_m^{TSLS} is identical to the regression objective function, just evaluated for $\tilde{x}$ instead of x. Thus, it has first-order-conditions given by $\rho\left(\theta^{TSLS}, y_i, \tilde{x}_i\right)$. Let $\kappa\left(\gamma,\tau,x_i^e,z_i,\omega_i\right) = E\left[\begin{pmatrix} z_i \\ \omega_i \end{pmatrix}\left(x_i^e-\gamma z_i-\tau\omega_i\right)\right]$ be the standard OLS moment conditions corresponding to (18) and let $\xi\left(\theta^{TSLS},\gamma,\tau,y_i,x_i^e,z_i,\omega_i\right) \equiv \xi\left(\theta^{TSLS},\gamma,\tau,\cdot\right)$ be the stacked vector of moment equations $\left(\ \kappa\left(\gamma,\tau,x_i^e,z_i,\omega_i\right)',\ \ \rho\left(\theta^{TSLS},y_{i,},\tilde{x}_i\right)'\ \right)'$. Denote the parameter vector combining both first and second stage coefficients as $v = \left(\ \theta^{TSLS\prime},\ \ \gamma',\ \ \tau',\ \right)'$. Define $\hat{v}$ as the estimated parameter vector solving the sample analogues of $\xi\left(\theta^{TSLS},\gamma,\tau,\cdot\right)$. As in the regression model, these moments constitute the basis for a GMM interpretation of the TSLS FCM model. Some additional assumptions are needed to characterize the asymptotic properties of $\hat{v}$. Denote as v^* the vector of true parameters from the first stage, γ_0, τ_0, and $\theta^{*,TSLS}$.

Assumption 6. *Additionally,*

1. $E\left[x_i^e z_i\right] < \infty$,$E\left[x_i^e w_i\right] < \infty$, $E\left[z_i^2\right] < \infty$,
2. v *is in the interior of* Υ; Υ *is compact.*
3. $\theta^{*,TSLS}$ *is unique,*
4. $\Gamma = E\left[\frac{\partial\xi(v^*,\cdot)}{\partial v'}\right]$ *is full rank,*
5. $E\left[\sup_{v\in\mathcal{N}}\left\|\frac{\partial\xi(v,\cdot)}{\partial v'}\right\|\right] < \infty$ *in a neighborhood* $\mathcal{N}$ *of* v^*,
6. $E\left[\xi\left(v,\cdot\right)\xi\left(v,\cdot\right)'\right]$ *is positive definite.*

Theorem 7. *Under Assumptions 5-2,*

1. $\hat{v}$, *is consistent for* v^*,

2. $\sqrt{N}\left(\hat{v}-v^{*}\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1} V^{TSLS} \Gamma'^{-1}\right)$, *where*

$$V^{TSLS} = E\left[\xi\left(v,\cdot\right)\xi\left(v,\cdot\right)'\right].$$

Proof. The proof largely follows from those of Theorems 5 and 6.For consistency, first note that γ_0 and τ_0 are unique solutions to the set of first-stage moment functions $\kappa\left(\gamma,\tau,x_i^e,z_i,\omega_i\right) = E\left[\begin{pmatrix} z_i \\ \omega_i \end{pmatrix}\left(x_i^e - \gamma z_i - \tau\omega_i\right)\right]$ corresponding to the OLS problem (18) by Assumption 5. v^* is thus the unique solution to $\xi\left(v\right)$. Since the second stage is just-identified (so the corresponding moments are always equal to zero at the optimum regardless of the value the first stage parameters take), the second stage has no influence on the first stage coefficients. It is immediate that $\kappa\left(v\right)$ is continuous in v. As noted in the proof of Corollary 4, the moment conditions $\rho\left(\theta, y_i, x_i\right)$ are continuous for all $\theta \in \Theta$. Since $\tilde{x}$ is continuous in γ,τ, this means that $\rho\left(\theta^{TSLS}, y_i, \tilde{x}_i\left(\gamma,\tau\right)\right)$ is continuous in v. The moments are bounded in expectation for all $v \in \Upsilon$ by duplicating the argument in the proof of Theorem 5 for the second stage under Assumption 5 and observing that the boundedness of the first-stage moments follows immediately from Assumption 6.1-2. These points jointly satisfy the requirements of standard GMM arguments, (e.g., Newey and McFadden (1994), p. 2121—2, Hayashi (2011) Proposition 7.7), so $\hat{v} \xrightarrow{p} v^*$.

For asymptotic normality, we first provide expressions for Γ to establish the continuous differentiability of $\xi\left(v\right)$ in v. Partition the blocks of Γ as

$$\Gamma = \begin{bmatrix} \Gamma_{\gamma\gamma} & \Gamma_{\gamma\tau} & \Gamma_{\gamma 1} & \cdots & \Gamma_{\gamma g} & \cdots & \Gamma_{\gamma G} \\ \Gamma_{\tau\gamma} & \Gamma_{\tau\tau} & \Gamma_{\tau 1} & \cdots & \Gamma_{\tau g} & \cdots & \Gamma_{\tau G} \\ \Gamma_{1\gamma} & \Gamma_{1\tau} & \Gamma_{11} & \cdots & \Gamma_{1g} & \cdots & \Gamma_{1G} \\ \vdots & \vdots & \vdots & \ddots & & & \vdots \\ \Gamma_{g\gamma} & \Gamma_{g\tau} & \Gamma_{g1} & & \Gamma_{gg} & & \Gamma_{gG} \\ \vdots & \vdots & \vdots & & & \ddots & \vdots \\ \Gamma_{G\gamma} & \Gamma_{G\tau} & \Gamma_{G1} & \cdots & \Gamma_{Gg} & \cdots & \Gamma_{GG} \end{bmatrix},$$

Note that since the moment conditions are no longer derived as the gradient of a single objective function, Γ is no longer a Hessian, and no longer symmetric, in particular the blocks linking the first and second stages. For the case where all second stage coefficients

are group-specific,

$$
\begin{aligned}
\Gamma_{\gamma\gamma} &= E\left[z_i^2\right] \\
\Gamma_{\tau\gamma} &= E\left[\omega_i z_i\right] \\
\Gamma_{\tau\tau} &= E\left[\omega_i \omega_i^{'}\right] \\
\Gamma_{\gamma g} &= 0 \\
\Gamma_{\tau g} &= 0 \\
\Gamma_{gg} &= E\left[\tilde{x}_i \tilde{x}_i' \left\{ \frac{-2m}{m-1} A_i^{-m-1} \left(e_{i,g}\right)^2 C_{i,g}^2 + \frac{m+1}{m-1} A_i^{-m} C_{i,g} \right\}\right] \\
\Gamma_{gh} &= E\left[\tilde{x}_i \tilde{x}_i' \left\{ \frac{-2m}{m-1} A_i^{-m-1} C_{i,h} e_{i,h} e_{i,g} C_{i,g} \right\}\right], h \neq g,
\end{aligned}
$$

with the elements of $\Gamma_{g\gamma}$ and $\Gamma_{g\tau}$ given by

$$\begin{aligned}
\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{g,1}^{TSLS}\partial\gamma} &= E\left[\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{h=1}^{G}\left(\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h}\theta_{h,1}^{TSLS}\right)C_{i,g}e_{i,g}\tilde{x}_i^e\right.\right.\\
&\left.\left.+A^{-m}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g}\theta_{g,1}^{TSLS}e_{i,g}\tilde{x}_i^e-A^{-m}C_{i,g}\theta_{g,1}^{TSLS}\tilde{x}_i^e+A^{-m}C_{i,g}e_{i,g}\right\}z_i\right]\\
&= E\left[A^{-m}C_{i,g}\left\{\frac{-2m}{m-1}A^{-1}\sum_{h=1}^{G}\left(C_{i,h}e_{i,h}\theta_{h,1}^{TSLS}\right)e_{i,g}\tilde{x}_i^e+\frac{m+1}{m-1}\theta_{g,1}^{TSLS}\tilde{x}_i^e+e_{i,g}\right\}z_i\right]\\
\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{g,k}^{TSLS}\partial\gamma} &= E\left[\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{h=1}^{G}\left(\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h}\theta_{h,1}^{TSLS}\right)C_{i,g}e_{i,g}\tilde{x}_{ik}\right.\right.\\
&\left.\left.+A^{-m}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g}\theta_{g,1}^{TSLS}e_{i,g}\tilde{x}_{ik}-A^{-m}C_{i,g}\theta_{g,1}^{TSLS}\tilde{x}_{ik}\right\}z_i\right]\\
&= E\left[A^{-m}C_{i,g}\tilde{x}_{ik}\left\{\frac{-2m}{m-1}A^{-1}\sum_{h=1}^{G}\left(C_{i,h}e_{i,h}\theta_{h,1}^{TSLS}\right)e_{i,g}+\frac{m+1}{m-1}\theta_{g,1}^{TSLS}\right\}z_i\right],k>1\\
\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{g,1}^{TSLS}\partial\tau'} &= E\left[\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{h=1}^{G}\left(\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h}\theta_{h,1}^{TSLS}\right)C_{i,g}e_{i,g}\tilde{x}_i^e\right.\right.\\
&\left.\left.+A^{-m}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g}\theta_{g,1}^{TSLS}e_{i,g}\tilde{x}_i^e-A^{-m}C_{i,g}\theta_{g,1}^{TSLS}\tilde{x}_i^e+A^{-m}C_{i,g}e_{i,g}\right\}\omega_i'\right]\\
&= E\left[A^{-m}C_{i,g}\left\{\frac{-2m}{m-1}A^{-1}\sum_{h=1}^{G}\left(C_{i,h}e_{i,h}\theta_{h,1}^{TSLS}\right)e_{i,g}\tilde{x}_i^e+\frac{m+1}{m-1}\theta_{g,1}^{TSLS}\tilde{x}_i^e+e_{i,g}\right\}\omega_i'\right]\\
\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{g,k}^{TSLS}\partial\tau'} &= E\left[\left\{\frac{-2m}{m-1}A^{-m-1}\sum_{h=1}^{G}\left(\left\|e_{i,h}\right\|^{-2m/(m-1)}e_{i,h}\theta_{h,1}^{TSLS}\right)C_{i,g}e_{i,g}\tilde{x}_{ik}\right.\right.\\
&\left.\left.+A^{-m}\frac{2m}{m-1}\left\|e_{i,g}\right\|^{\frac{-4m+2}{m-1}}e_{i,g}\theta_{g,1}^{TSLS}e_{i,g}\tilde{x}_{ik}-A^{-m}C_{i,g}\theta_{g,1}^{TSLS}\tilde{x}_{ik}\right\}\omega_i'\right]\\
&= E\left[A^{-m}C_{i,g}\tilde{x}_{ik}\left\{\frac{-2m}{m-1}A^{-1}\sum_{h=1}^{G}\left(C_{i,h}e_{i,h}\theta_{h,1}^{TSLS}\right)e_{i,g}+\frac{m+1}{m-1}\theta_{g,1}^{TSLS}\right\}\omega_i'\right],k>1
\end{aligned}$$

where $e_{i,g}=y_i-\theta_g^{TSLS}\tilde{x}_i$, $A_i=\Sigma_{g=1}^{G}\left\|e_{i,g}\right\|^{-2/(m-1)}$, $C_{ig}=\left\|e_{i,g}\right\|^{-2m/(m-1)}$ and we have exploited the fact that $\left\|e_{i,g}\right\|^2=e_{i,g}^2$ since $T=1$. We also provide expressions for elements of Γ that change when there are controls in ω_k with common coefficients across groups,

such that $\theta_{g,k}^{TSLS} = \theta_{h,k}^{TSLS} \equiv \theta_{\star,k}^{TSLS}, h \neq g$. In this case,

$$\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{\star,k}^{TSLS}\partial\theta_{\star,k}^{TSLS}} = E\left[\tilde{x}_{i,k}^2 A_i^{-m}\left\{\frac{-2m}{m-1}A_i^{-1}B_i^2 + \frac{m+1}{m-1}\sum_{g=1}^{G} C_{i,g}\right\}\right]$$
$$\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{\star,k}^{TSLS}\partial\theta_{\star,l}^{TSLS}} = E\left[\tilde{x}_{i,k}\tilde{x}_{i,l} A_i^{-m}\left\{\frac{-2m}{m-1}A_i^{-1}B_i^2 + \frac{m+1}{m-1}\sum_{g=1}^{G} C_{i,g}\right\}\right]$$
$$\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{\star,k}^{TSLS}\partial\theta_{g,l}^{TSLS}} = E\left[\tilde{x}_{i,k}\tilde{x}_{i,l} A_i^{-m}\left\{\frac{-2m}{m-1}A_i^{-1}C_{i,g}e_{i,g}B_i + \frac{m+1}{m-1}C_{i,g}\right\}\right],$$
$$\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{\star,k}^{TSLS}\partial\gamma} = E\left[\tilde{x}_{i,k} z_i A_i^{-m}\left\{\frac{-2m}{m-1}A_i^{-1}\sum_{g=1}^{G}\left(C_{i,g}e_{i,g}\theta_{g,1}^{TSLS}\right)B_i + \frac{m+1}{m-1}\sum_{g=1}^{G} C_{i,g}\theta_{g,1}^{TSLS}\right\}\right]$$
$$\frac{\partial^2 L_m^{TSLS}}{\partial\theta_{\star,k}^{TSLS}\partial\tau'} = E\left[\tilde{x}_{i,k} A_i^{-m}\left\{\frac{-2m}{m-1}A_i^{-1}\sum_{g=1}^{G}\left(C_{i,g}e_{i,g}\theta_{g,1}^{TSLS}\right)B_i + \frac{m+1}{m-1}\sum_{g=1}^{G} C_{i,g}\theta_{g,1}^{TSLS}\right\}\omega_i'\right]$$

where $B_i = \sum_{g=1}^{G}\left[e_{i,g}C_{i,g}\right]$. By inspection, Γ is continuous in v, since $e_{i,g}, A_i^{-m}, A_i^{-m-1}, C_{ig}, B_i$ are continuous in θ^{TSLS}, and all elements of Γ are continuous functions of these objects.

Next, we establish the asymptotic normality of $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi\left(v\right)$. Since $y_i, x_i^e, \omega_i, z_i$ are assumed to be jointly i.i.d., $\xi\left(v, y_i, x_i^e, \omega_i, z_i\right)$ is i.i.d. across observations, so by the Lindeberg-Levy central limit theorem,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\xi\left(v,\cdot\right) \xrightarrow{d} \mathcal{N}\left(0, V^{TSLS}\right),$$

where $V^{TSLS} = E\left[\xi\left(v,\cdot\right)\xi\left(v,\cdot\right)'\right]$ is assumed to be positive definite in Assumption 6.6.

Combining these two results with the additional conditions of Assumption 6, the standard conditions for asymptotic normality of a GMM estimator are satisfied (e.g.,Hayashi (2011) Proposition 7.10). Since the weighting matrix is the identity (we assumed the problem is just-identified),

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}V^{TSLS}\Gamma'^{-1}\right).$$

□

B.3 Fixed$-T$ asymptotic variance for HKM with heterogeneous coefficients

Here we report fixed-T analytical formulas that extend those reported in the Appendix of Bonhomme and Manresa (2015) to the case of heterogeneous slope coefficients. We adjust notation slightly from the remainder of our paper to be consistent with Bonhomme and Manresa (2015), using α_g for fixed effects, θ_g for group-specific coefficients, and β for common coefficients; however, we limit our attention to the $T = 1$ and univariate x case given our empirical focus.

The objective function given estimated groups $\hat{g}_j(\theta, \alpha, \beta)$ takes the form

$$E\left[\left(y_j - \theta_{\hat{g}_j(\theta,\alpha,\beta)} x_j - \alpha_{\hat{g}_j(\theta,\alpha,\beta)} - W_j\beta\right)^2\right],$$

which yields the moment equations

$$\begin{aligned} E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} x_j \left(y_j - \bar{\theta}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} x_j - \bar{\alpha}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} - W_j'\bar{\beta}\right)\right] &= 0, \\ E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} \left(y_j - \bar{\theta}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} x_j - \bar{\alpha}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} - W_j'\bar{\beta}\right)\right] &= 0, \\ E\left[W_j \left(y_j - \bar{\theta}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} x_j - \bar{\alpha}_{\hat{g}_j(\bar{\theta},\bar{\alpha},\bar{\beta})} - W_j'\bar{\beta}\right)\right] &= 0, \end{aligned}$$

for the solution $\left(\bar{\theta}, \bar{\alpha}, \bar{\beta}\right)$. Thus, the Jacobian of the moment conditions has the form

$$\Gamma = \begin{pmatrix} \Gamma_{\beta\beta} & \Gamma_{\beta\theta_1} & \dots & \Gamma_{\beta\theta_G} & \Gamma_{\beta\alpha_1} & \dots & \Gamma_{\beta\alpha_G} \\ \Gamma_{\theta_1\beta} & \Gamma_{\theta_1\theta_1} & \dots & \Gamma_{\theta_1\theta_G} & \Gamma_{\theta_1\alpha_1} & \dots & \Gamma_{\theta_1\alpha_G} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Gamma_{\theta_G\beta} & \Gamma_{\theta_G\theta_1} & \dots & \Gamma_{\theta_G\theta_G} & \Gamma_{\theta_G\alpha_1} & \dots & \Gamma_{\theta_G\alpha_G} \\ \Gamma_{\alpha_1\beta} & \Gamma_{\alpha_1\theta_1} & \dots & \Gamma_{\alpha_1\theta_G} & \Gamma_{\alpha_1\alpha_1} & \dots & \Gamma_{\alpha_1\alpha G} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \Gamma_{\alpha_G\beta} & \Gamma_{\alpha_G\theta_1} & \dots & \Gamma_{\alpha_G\theta_G} & \Gamma_{\alpha_G\alpha_1} & \dots & \Gamma_{\alpha_G\alpha_G} \end{pmatrix},$$

where notation follows Bonhomme and Manresa (2015) which means there are 9 unique elements to characterize. Explicit expressions are given below.

$$\Gamma_{\beta\beta} = E\left[W_j W_j'\right] + \sum_{g=1}^{G}\sum_{h\neq g} E\left[\left(\int_{\bar{S}_{gh}} f\left(y \mid x_j\right) dy\right)\left(\bar{\theta}_g x_j + \bar{\alpha}_g\right)\frac{\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} W_j W_j'\right],$$

$$\Gamma_{\beta\theta_g} = E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} W_j x_j\right]$$
$$+ \sum_{h\neq g} E\left[W_j\left(\bar{\theta}_g x_j + \bar{\alpha}_g - \bar{\theta}_h x_j - \bar{\alpha}_h\right)\left(\int_{\bar{S}_{gh}} \frac{x_j\left(y - x\bar{\theta}_g - \bar{\alpha}_g - W_j'\bar{\beta}\right)}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right)\right],$$

$$\Gamma_{\beta\alpha_g} = E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} W_j\right]$$
$$+ \sum_{h\neq g} E\left[W_j\left(\bar{\theta}_g x_j + \bar{\alpha}_g - \bar{\theta}_h x_j - \bar{\alpha}_h\right)\left(\int_{\bar{S}_{gh}} \frac{y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right)\right],$$

$$\Gamma_{\theta_g\theta_g} = E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} x_j^2\right] - E\left[\sum_{h\neq g}\left(\int_{\bar{S}_{gh}} \frac{x_j^2\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)^2}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right)\right],$$

$$\Gamma_{\theta_g\theta_{\tilde{g}}} = E\left[\int_{\bar{S}_{gh}} \frac{x_j^2\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)\left(y - \bar{\theta}_{\tilde{g}} x_j - \bar{\alpha}_{\tilde{g}} - W_j'\bar{\beta}\right)}{\left\|\bar{\theta}_{\tilde{g}} x_j + \bar{\alpha}_{\tilde{g}} - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right],$$

$$\Gamma_{\theta_g\alpha_g} = E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\} x_j\right] - E\left[\sum_{h\neq g}\left(\int_{\bar{S}_{gh}} \frac{x_j\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)^2}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right)\right],$$

$$\Gamma_{\theta_g\alpha_{\tilde{g}}} = E\left[\int_{\bar{S}_{gh}} \frac{x_j\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)\left(y - \bar{\theta}_{\tilde{g}} x_j - \bar{\alpha}_{\tilde{g}} - W_j'\bar{\beta}\right)}{\left\|\bar{\theta}_{\tilde{g}} x_j + \bar{\alpha}_{\tilde{g}} - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right],$$

$$\Gamma_{\alpha_g\alpha_g} = E\left[\mathbf{1}\left\{\hat{g}_j\left(\bar{\theta},\bar{\alpha},\bar{\beta}\right) = g\right\}\right] - E\left[\sum_{h\neq g}\left(\int_{\bar{S}_{gh}} \frac{\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)^2}{\left\|\bar{\theta}_h x_j + \bar{\alpha}_h - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right)\right],$$

$$\Gamma_{\alpha_g\alpha_{\tilde{g}}} = E\left[\int_{\bar{S}_{gh}} \frac{\left(y - \bar{\theta}_g x_j - \bar{\alpha}_g - W_j'\bar{\beta}\right)\left(y - \bar{\theta}_{\tilde{g}} x_j - \bar{\alpha}_{\tilde{g}} - W_j'\bar{\beta}\right)}{\left\|\bar{\theta}_{\tilde{g}} x_j + \bar{\alpha}_{\tilde{g}} - \bar{\theta}_g x_j - \bar{\alpha}_g\right\|} f\left(y \mid x_j\right) dy\right].$$

Estimators can easily be constructed for these analytical expressions using the kernel approach of Bonhomme and Manresa (2015). The natural estimator for the moment condition covariance takes the form

$$\hat{V} = \frac{1}{N}\sum_{j=1}^{N}\begin{pmatrix} x_j \iota_{\hat{g}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right)} \\ \iota_{\hat{g}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right)} \\ W_j \end{pmatrix} \hat{v}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right)^2 \begin{pmatrix} x_j \iota_{\hat{g}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right)} \\ \iota_{\hat{g}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right)} \\ W_j \end{pmatrix}'$$

for residuals $\hat{v}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right) \equiv \hat{v}_j\left(\hat{g}_j\left(\hat{\theta},\hat{\alpha},\hat{\beta}\right),\hat{\theta},\hat{\alpha},\hat{\beta}\right)$.

B.4 Inclusion of controls in the objective function

In this section, we highlight the role that controls play in our model. In particular, we show that the baseline specification in Equation (19) without controls included can be seen as a non-parametric alternative to a specification with controls as the number of groups, G, increases.

Consider two models. Suppose the true model has the form

$$\Delta C_j = \beta' W_j + \sum_{g \in G} \left(\theta_g \mathbf{1}\left[j \in g\right] R_j + \alpha_g \mathbf{1}\left[j \in g\right] \right) + \epsilon_j. \tag{19}$$

Second, suppose the econometrician estimates a simpler model, omitting the controls:

$$\Delta C_j = \sum_{\tilde{g} \in \tilde{G}} \left(\tilde{\theta}_{\tilde{g}} \mathbf{1}\left[j \in \tilde{g}\right] R_j + \tilde{\alpha}_{\tilde{g}} \mathbf{1}\left[j \in \tilde{g}\right] \right) + \tilde{\epsilon}_j. \tag{20}$$

In general, this second model could be susceptible to omitted variable bias (in particular, if $\tilde{G} = G$). However, this need not be the case if $\tilde{G}$ is allowed to vary. This is because the model in Equation (19) can be rewritten in the form of Equation (20), with $\tilde{G} \geq G$. To see this, consider a simple example where W_j is a scalar binary regressor and $G = 2$. Then if $\tilde{G} = 4$, there are four cases to consider, based on two "true" groups with heterogeneous parameters, and two levels of W_j within each group. Then the following relationships exist between $\left\{\theta_g, \alpha_g\right\}_{g=1,2}$ and $\left\{\tilde{\theta}_{\tilde{g}}, \tilde{\alpha}_{\tilde{g}}\right\}_{\tilde{g}=1,\ldots,4}$ (where the labels of groups are arbitrary):

expanded group	true group	control value	slope	intercept
$\tilde{g} = 1$	$g = 1$	$W_j = 0$	$\tilde{\theta}_1 = \theta_1$	$\tilde{\alpha}_1 = \alpha_1$
$\tilde{g} = 2$	$g = 1$	$W_j = 1$	$\tilde{\theta}_2 = \theta_1$	$\tilde{\alpha}_2 = \alpha_1 + \beta$
$\tilde{g} = 3$	$g = 2$	$W_j = 0$	$\tilde{\theta}_3 = \theta_2$	$\tilde{\alpha}_3 = \alpha_2$
$\tilde{g} = 4$	$g = 2$	$W_j = 1$	$\tilde{\theta}_3 = \theta_2$	$\tilde{\alpha}_4 = \alpha_2 + \beta$

The effect of W_j is absorbed entirely into the fixed effects $\tilde{\alpha}_{\tilde{g}}$, which now vary with an individual's W_j. The true values of θ_g and α_g are still recovered, provided $\tilde{G}$ is chosen correctly.

This argument can be extended to allow for effects of W_j other than simple level shifts. For example, if the true model has the additional interaction term $\theta_g^W \mathbf{1}\left[j \in g\right] R_j W_j$, then the $\tilde{\theta}_{\tilde{g}}$'s recovered would incorporate θ_g^W just like the expressions for $\tilde{\alpha}_{\tilde{g}}$ above incorporate

β. The argument also extends to non-binary controls. For example, a discrete regressor taking k values would expand a $G-$group model to a $k \times G$ group model. Admittedly, extending the argument to a continuous regressor introduces a computational challenge in practice, but in our setting, available controls are generally discrete. Finally, the argument generalizes in the same way when W_j is a vector and not a simple scalar.

There remains a question over whether estimating such a model accurately recovers the heterogeneity in θ_g, or rather overestimates heterogeneity as G grows to $\tilde{G}$. In our view, $\tilde{G}$ more accurately represents the true heterogeneity in the underlying data, seeing as it incorporates any differences in MPCs arising from observable controls, W_j, as opposed to only residual heterogeneity after partialing out W_j. Individuals still have meaningfully different MPCs, even if that difference is explained by observable characteristics. We can then, of course, investigate the relationship between the recovered $\tilde{\theta}_{\tilde{g}}$ and W_j *ex post*, as we do in Section 5.3. We seek to characterize the full heterogeneity of MPCs, as opposed to the conditional heterogeneity of MPCs, as for instance in Kaplan et al. (2014), Fagereng et al. (2016), Johnson et al. (2006), Parker et al. (2013), and Crawley and Kuchler (2018).

A further advantage of estimating Equation (20) as opposed to (19) is that it allows the relationship between W_j and ΔC_j to be completely non-parametric. Including W_j as in Equation (19) assumes the term enters linearly; including an interaction with $\mathbf{1}\left[j \in g\right] R_j$ likewise assumes a functional form. However, estimating a separate set of parameters $\tilde{\theta}_{\tilde{g}}, \tilde{\alpha}_{\tilde{g}}$ for each $\tilde{g} \in \tilde{G}$ takes no stance on the parametric structure relating W_j to C_j. On this basis, as well as the desire to recover the full heterogeneity in MPCs, we proceed using specifications based on Equation (20) as our baseline.

These insights inform our empirical specification in Equation (14). In particular, we opt to include a minimal set of covariates, W_j, including time dummies, age, age squared, and changes in household membership, and explore the relationship between additional covariates and consumption behavior in Section 5.3.

C Supplemental numerical and simulation results

C.1 Numerical results for optimal m in Gaussian cluster means

In this section we consider the cluster means case of section 2.1. In figure 10 we show how the estimated positive mean by FCM, as we change the fuzziness parameter m. As documented analytically, there is a $m = \tilde{m}$ such that bias is zero. We report these $\tilde{m}$ in table 1.

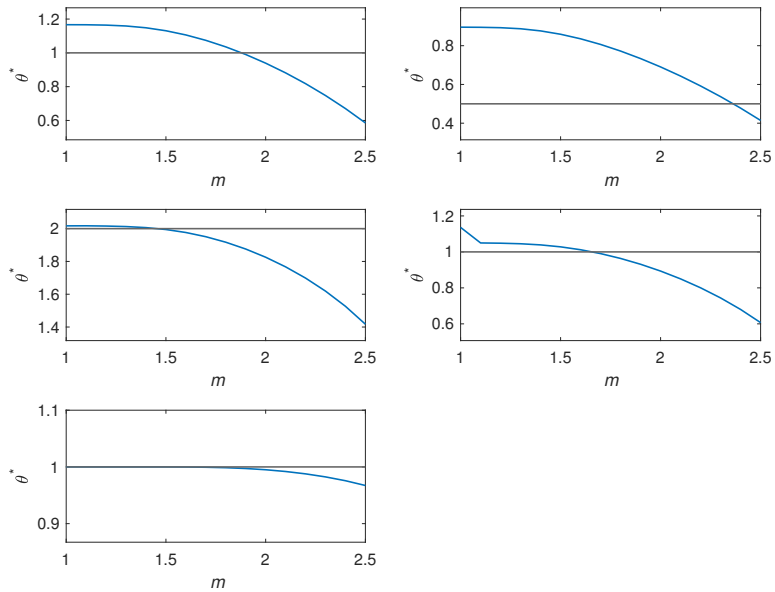


Figure 10: Numerical results for FCM estimated means as a function of m.

C.2 Computational performance

Our approach entails two computational improvements. First, thanks to the equivalence shown in Theorem 2, we are able to improve vis a vis an iterative FCM approach. Second, we greatly improve computational speed with respect to various HKM algorithms.

The structure of our problem is such that we can apply standard non-linear minimization routines to the objective function, and make use of analytical gradients to further enhance performance. This strategy is substantially faster than an iterative procedure aimed at achieving the minimizer weights via convergence algorithms. The improvements in computational speed increase with the chosen number of groups. With 10 groups, for instance, our GMM approach is more than twice as fast as its iterative counterpart.[37]

The FCM algorithm is also faster than HKM. In the simulations shown in Section 3.2, we consider two versions of the algorithm proposed by Bonhomme and Manresa (2015). "Algorithm 1", directly from that paper, starts by randomly allocating observations to a given number of groups, estimates the model, and iteratively reallocates observations while estimating the model until convergence. This strategy is considerably slower than FCM; moreover, it performs much more poorly in simulations, as we show later. It takes about 6 seconds to solve our FCM algorithm with 5 groups, for one vector of initial guesses. We show that 500 starting guesses are more than enough to obtain consistent and stable results. Algorithm 1 takes instead between 5 and 9 minutes, over the same

[37] We solve the model, for one vector of initial coefficients, in 17 seconds, compared with 40 seconds required to solve the model iteratively.

machine (MacBook Pro, 2.7Ghz Intel Core with turbo boost up to 3.8Ghz, 16GB memory), with the same number of starting values. Moreover, we find that even 1,000 starting values are not enough to ensure stable results.

Bonhomme and Manresa (2015) propose a variable neighborhood search algorithm, "Algorithm 2", which, in their framework, improves simulation performance and is often necessary to avoid local minima. In our relatively large cross-sectional dataset, this algorithm proves infeasible.[38] We therefore consider an intermediate version, which we label "Algorithm 1.5". This uses the best result of Algorithm 1 (with 500 starting values) as the start value of a variable neighborhood approach, which however excludes the local search component. This algorithm repeats the assignment procedure relocating n randomly selected individuals and iteratively proceeds until the objective function stops improving. It increases $n$by 1 until n_{max}, until the objective decreases. When this happens, it restarts $n$and repeats the procedure for j outer iterations. Jumps of size n allow escape from local solutions trapped in valleys. In the simulations shown in section 3.2, we set both n_{max} and j to 100. We also set a stopping rule of 30 iterations which exits the algorithm if 30 consecutive j iterations do not improve the objective function.[39] Within every variable neighborhood iteration, Algorithm 2 systematically checks all re-assignments of individual observations across groups, updating group assignment when the objective function decreases. This step took more than 1 week to perform for only 1 starting value in our dataset. Algorithm 1.5, instead, adds only a further 9 minutes to Algorithm 1. Its nature, however, does not allow parallelization beyond the one on starting guesses used for Algorithm 1.

C.3 Additional simulation results

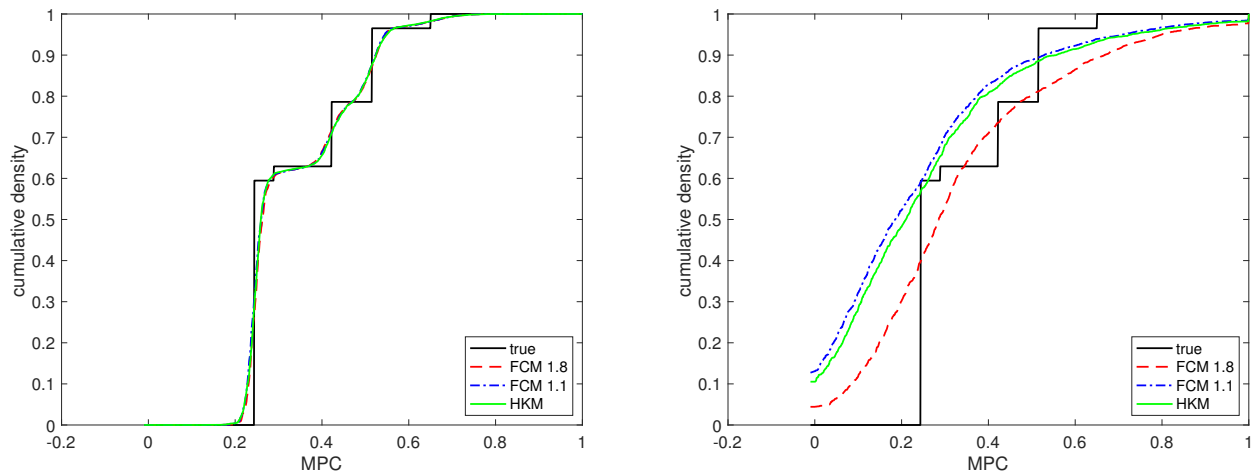
In this section we complement the simulation results shown in section 3.2. We make use of empirical CDFs to show that FCM performance has remarkable advantages not only when looking at MPC point estimates, but also considering the whole MPC distribution. We start, in figure 12a, from the simulation using Gaussian errors with empirical noise, as in Table 1 . For each sample, we compute the empirical CDF of the estimated distribution of modal MPCs, over a fixed grid bounded between 0 and 1.[40] We then report the average of these CDFs across samples. All models do very well in matching the true CDF. Aver-

[38]The infeasibility stems from the fact that one step of the algorithm - local search - requires looping over all observations (~17K in our case) sequentially in every iteration, which cannot be parallelized.

[39]In appendix C.3 we increase those parameters and show that HKM performance improves only mildly, at the expenses of remarkable increases in computational time.

[40]For FCM, we could also look the weighted MPC. Results are broadly unchanged after averaging across samples.

Figure 11: Empirical CDFs of the MPC distribution: G = 5



(a) Gaussian errors, empirical noise (b) Empirical errors

Notes: 500 samples, generated as in section 3.2. Averages across samples of the empirical CDF of modal MPC distributions.

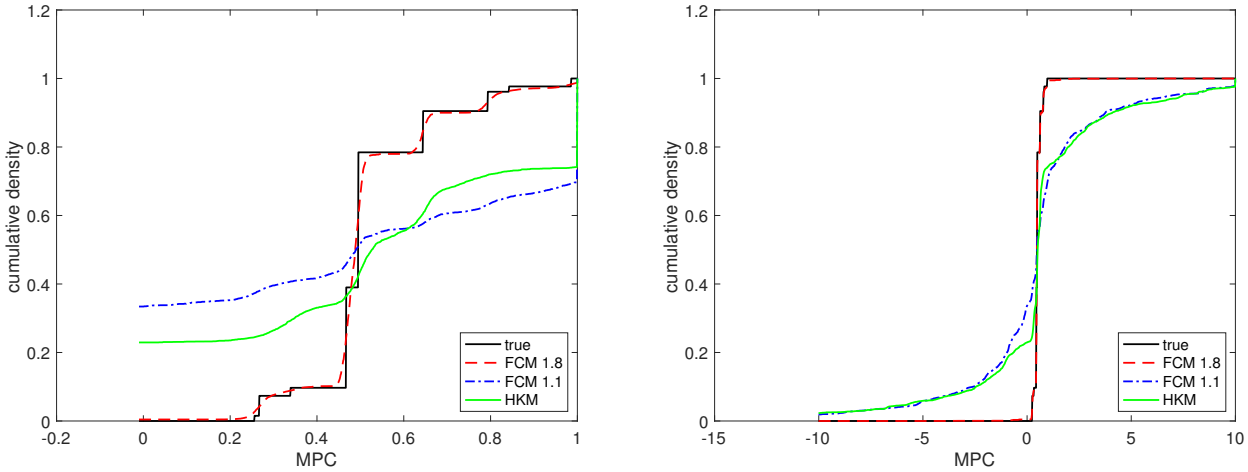
aging smooths the CDF, taking into account sample uncertainty as shown by the RMSE. The median CDF almost perfectly replicates the DGP.

We then turn to the data simulated using empirical errors, as in Table 2. While the fit is less good, our baseline model clearly does a better job in fitting the true distribution. Since some of the estimated MPCs are beyond the fixed domain over which we evaluate the CDF, there is some non-negligible mass at the left end of the distribution, especially for HKM and FCM with small m.

In Table 3 we have shown how our benchmark performs better than the alternatives, both in terms of point estimates and RMSE. We corroborate these findings showing the empirical CDFs. When estimating the CDF over a narrow grid, between 0 and 1, HKM predicts that almost 50% of the observations will have MPCs beyond those boundaries. Our benchmark model, instead, does a strikingly good job at matching the data, as shown in figure 13a. When we extend the grid to encompass most of the estimated MPCs in HKM, it is even more clear the extent to which some MPCs are wrongly estimated, and the portion of the population for which it matters.

Finally, we have discussed in section C.2 how HKM is a computationally intensive algorithm, especially when the number of groups increases. We increase all the tuning parameters in order to explore whether this delivers improvements in simulation performance, albeit implying a remarkable increase in computational cost. In particular, we increase the starting values in Algo. 1 to 2,000. Moreover, we increase j to 150, n_{max} to 200, and the stopping rule to 100, see C.2. In Table 5 we show that some point estimates

Figure 12: Empirical CDFs of the MPC distribution: G = 10



(a) Gaussian errors, empirical noise, narrow grid

(b) Gaussian errors, empirical noise, full grid

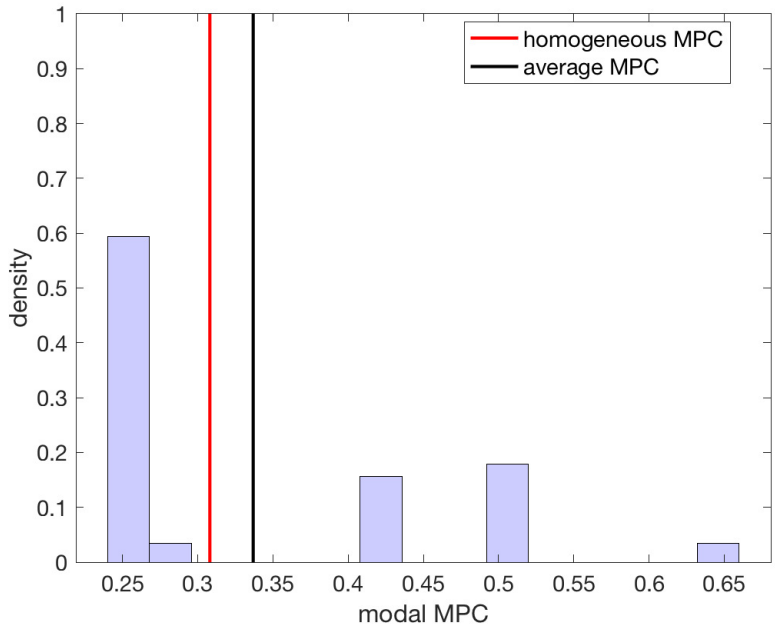
Notes: 500 samples, generated as in section 3.2. Averages across samples of the empirical CDF of modal MPC distributions.

get closer to the truth, and some RMSE fall, although the performance of our benchmark FCM model seems to still be unambiguously better.

	Truth	FCM		HKM	
		$m = 1.8$	$m = 1.1$	Algo. 1	Algo. 1.5
Point Estimates	0.844	0.784	0.668	-6.797	2.320
	0.986	1.032	2.136	2.636	16.766
	0.795	0.797	1.049	1.610	1.684
	0.646	0.646	0.499	0.715	0.410
	0.496	0.495	0.456	0.703	0.535
	0.468	0.477	0.460	0.705	0.745
	0.496	0.488	0.401	0.469	0.598
	0.268	0.269	0.354	-1.052	0.351
	0.340	0.344	0.443	-2.839	0.276
	0.257	0.263	0.505	16.448	0.545
RMSE		0.676	5.932	11.313	6.600
		0.771	17.218	5.173	18.199
		0.029	3.829	3.347	3.494
		0.023	2.530	3.170	2.275
		0.094	1.599	2.346	0.906
		0.129	3.547	2.315	1.206
		0.174	1.769	4.182	1.317
		0.023	0.389	3.917	0.298
		0.048	0.910	5.259	0.993
		0.056	0.421	18.095	0.381
Rejection Rates		0.088	0.758	0.824	0.852
		0.106	0.928	0.972	0.988
		0.074	0.904	0.848	0.928
		0.086	0.892	0.892	0.704
		0.066	0.882	0.700	0.524
		0.058	0.880	0.560	0.632
		0.060	0.754	0.832	0.524
		0.054	0.456	0.776	0.372
		0.070	0.610	0.860	0.892
		0.058	0.508	0.988	0.668
Share Misclassified		0.007	0.409	0.958	0.660

Table 5: Simulation, Gaussian errors, empirical noise, $G^* = 10$, $S = 250$

Figure 13: Estimated distribution of MPCs out of the tax rebate: modal MPC



Notes: Figure 13 plots a histogram of estimated MPCs out of total expenditures, defined as in Parker et al. (2013). The homogeneous MPC (red line) is estimated assuming homogeneous response to the tax rebate, as in Parker et al. (2013). The black line shows the average modal MPC in our sample.

D Supplemental empirical results

D.1 The MPC distribution: additional results

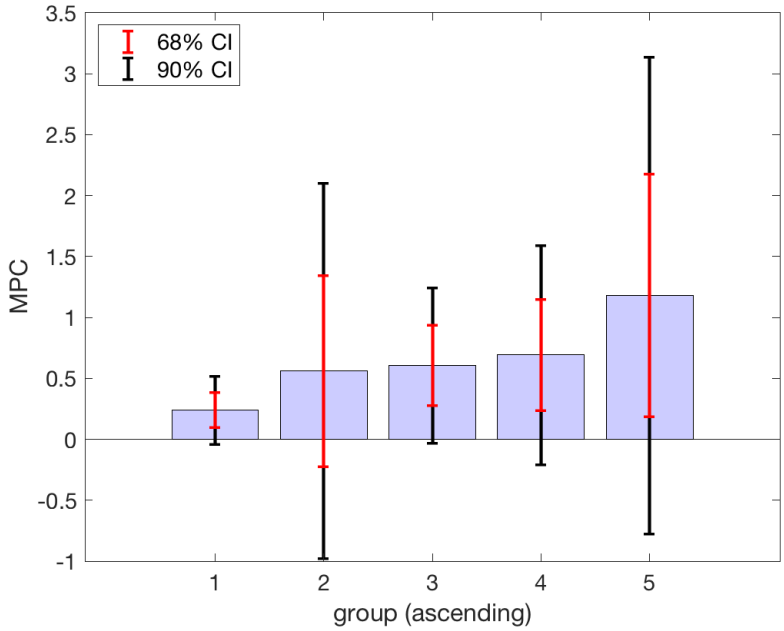
Figure 13 shows the distribution of modal MPCs out of total expenditures.

Figure 14 reports the 68% and 90% confidence bands of our estimated MPCs out of total expenditures, estimated via TSLS.

Moreover, we further confirm the reliability of our estimated MPC distribution as follows. We draw 100 samples via bootstrap with replacement. For each sample, we estimate our FCM algorithm (using the baseline specification). Table 6 shows how, on average, the bootstrapped samples generate quantiles of the weighted MPC distribution that are very close to those of the empirical distribution shown in Figure 2. This shows that even though MPC estimates may not be individually statistically significant, there is relatively little sampling uncertainty surrounding the overall shape of the distribution.

We also show in table 7 that nearly all the nondurable MPCs are statistically different from each other when we estimate a heterogeneous WLS taking weights as given, as explained in Table 4.

Figure 14: Estimated MPCs out of the Tax Rebate: TSLS



Notes: Figure 14 depicts the estimated MPCs for total expenditures (defined as in Parker et al. (2013)) each of the $G = 5$ groups using the TSLS specification. 90% confidence intervals (black lines) and 68% confidence intervals (red lines) are depicted in vertical lines. The confidence intervals are constructed using the analytical formulas derived in Theorem 6.

Table 6: Average quantiles of the MPC distribution across bootstrapped samples

	Average	p10	p25	p50	p75	p90
Data	0.341	0.245	0.249	0.293	0.422	0.506
Bootstrap	0.357	0.231	0.247	0.292	0.433	0.570
	(0.142)	(0.073)	(0.121)	(0.102)	(0.116)	(0.169)

Notes: Table 6 shows various statistics of the distribution of weighted MPCs in the data (first row) and in a bootstrap exercise. pxx, etc. denotes the xxth percentiles. "Data" refers to Figure 2, while the row labeled "Bootstrap" shows the average of each moment across 100 bootstrapped samples with replacement. The last rows the standard deviation of each moment across bootstrapped samples.

Table 7: Test for MPC equality: non-durables

	MPC				
	0.01	0.08	0.15	0.18	0.33
0.01	14.9				
	(0.00)				
0.08	12.8	0.23			
	(0.00)	(0.63)			
0.15	4.28	22.0	64.7		
	(0.04)	(0.00)	(0.00)		
0.18	0.87	44.6	10.6	86.7	
	(0.35)	(0.00)	(0.00)	(0.00)	
0.33	4.88	52.4	31.8	18.0	60.5
	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)

Notes: The table shows F-statistics from pairwise two-sided Wald tests of equality across MPCs (the diagonals shows tests of equality with zero). Weights are taken as given. P-values are reported in parentheses.

Table 8: Individual correlations with the MPC out of total expenditures

	Log salary income	Log total income	Mortgage interest to income ratio	APC	Age	Log liquid wealth
OLS weighted MPC	0.13***	0.20***	0.08***	0.11***	-0.06***	0.12***
TSLS weighted MPC	0.13***	0.21***	0.06***	0.15***	-0.06***	0.13***

Notes: Table 8 shows the correlations between estimates listed in rows and observables listed in columns. *, ** and *** denote significance of the correlation at 10, 5 and 1% respectively.

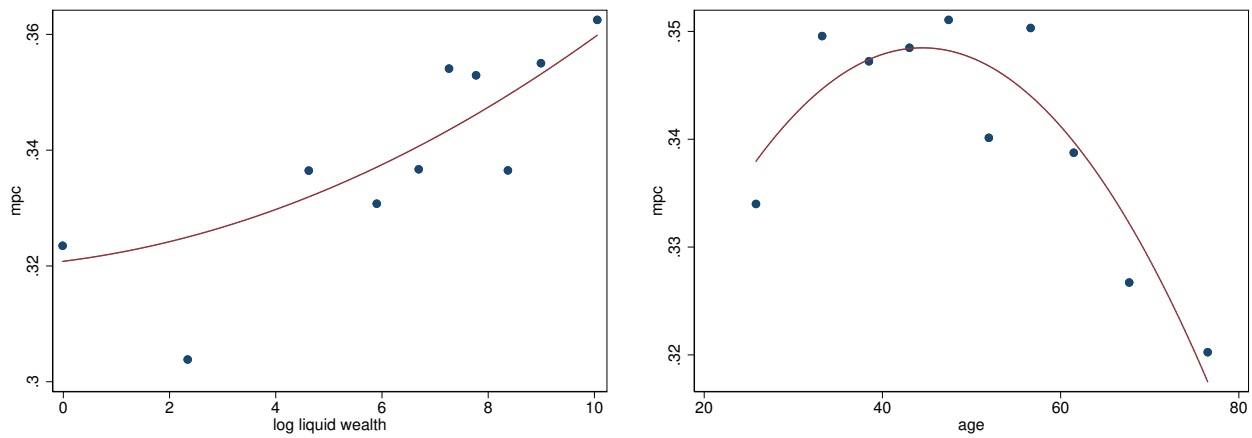
D.2 What drives MPC heterogeneity: additional results

Some household characteristics individually correlate with the MPC distribution, although this correlation breaks down or becomes insignificant when considering additional controls. Table 8 shows the individual correlations between a set of observables and the MPC.

We then turn to analyze whether the linear correlation with age and liquid wealth hides some non-linear pattern. Figure 16a suggests a positive and convex relationship between the weighted MPC and log liquid wealth. The relationship looks instead concave with respect to age, as shown in Figure 16b.

Neither relationship is robust to the inclusion of a set of controls. In table 9 we show that the best array of observable predictors explains only 13% of the variance in weighted

Figure 15: Marginal propensities to consume: liquid wealth and age



(a) Liquid wealth

(b) Age

Notes: Binscatter. Each dot shows the average weighted MPC out of total expenditures for each decile of the distribution of lagged log liquid wealth (left panel) and for each decile of the distribution of age of the reference person in the household (right panel). Log of liquid wealth takes 0 when liquid wealth is 0 or negative.

MPC. Moreover, we highlight how the relationship between the MPC distribution and observable predictors is broadly unaltered when considered MPCs estimated via TSLS. The same, is true, although to a lesser extent, for nondurable and durable MPCs, as shown in table 10.

Finally, we explore nonlinear effects of observable predictors of the MPC distribution. We estimate a multinomial logit model, using the modal MPC as the dependent variable. Table 11 shows the estimation output, whereas Figure 16 plots the marginal effects for the APC.[41] The results confirm that total income, APC, and mortage are the three main correlates with the MPC distribution, even when we allow for nonlinear effects.

D.3 Rebate coefficient versus MPC

Following Kaplan and Violante (2014), we modify specification 13 by introducing the lag of the rebate variable R_j^{lag} so that the estimated rebate coefficient can be interpreted as an MPC:

$$\Delta C_j = \beta' W_j + \theta R_j + \theta^{\text{lag}} R_j^{\text{lag}} + \alpha + \epsilon_j \tag{21}$$

[41] We report only the results for the MPC distribution estimated via OLS and an array of predictors that excludes liquid wealth. Results are basically unchanged if we look at 2SLS MPC or include liquidity.

Table 9: Explanatory variables: weighted MPC out of total expenditures

	OLS weighted MPC		TSLS weighted MPC	
	(I)	(II)	(III)	(IV)
Log salary income	0.002	-0.0003	0.003	0.001
	(0.02)	(0.001)	(0.030)	(0.002)
Log total income	0.036***	0.045***	0.081***	0.095***
	(0.008)	(0.006)	(0.015)	(0.012)
Mortgage interest to income ratio	0.086***	0.060**	0.141**	0.087*
	(0.032)	(0.027)	(0.061)	(0.052)
APC	0.044***	0.053***	0.111***	0.127***
	(0.008)	(0.007)	(0.016)	(0.013)
Outright homeowner dummy	0.006	0.008	0.006	0.009
	(0.011)	(0.010)	(0.021)	(0.019)
Mortgagor dummy	0.029***	0.023**	0.045**	0.035*
	(0.011)	(0.010)	(0.022)	(0.018)
Age	-0.002	-0.001	-0.004	-0.004
	(0.002)	(0.001)	(0.003)	(0.003)
Age-squared	0.000	0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Married dummy	0.011	0.011	0.032**	0.028**
	(0.008)	(0.007)	(0.015)	(0.013)
Number of children	0.006	0.0001	0.006	-0.002
	(0.006)	(0.003)	(0.007)	(0.006)
Log liquid wealth	-0.001		-0.001	
	(0.001)		(0.002)	
R^2	0.13	0.13	0.16	0.15
Number of observations	723	1,079	723	1,079

Notes: All logged variables takes 0 when the raw value is 0 or negative. *, ** and *** denote significance of the coefficients at 10, 5 and 1% respectively. Standard errors in parentheses. Age and its square are controls in our FCM estimation. While this does not pose an issue for the point estimates shown in this table, it might affect inference. We repeated the same regressions shown here, excluding age and age-sq, and all the coefficients were unaffected.

Table 10: Explanatory variables: weighted MPC out of nondurables and durables

	Nondurables weighted MPC		Durables MPC (dummy)	
	(I)	(II)	(III)	(IV)
Log salary income	0.001	-0.002	0.005	0.004
	(0.001)	(0.001)	(0.004)	(0.003)
Log total income	0.015**	0.021***	-0.002	0.009
	(0.006)	(0.005)	(0.022)	(0.016)
Mortgage interest to income ratio	0.006	0.012	0.103	0.085
	(0.025)	(0.021)	(0.091)	(0.071)
APC	0.017***	0.019***	-0.002	0.003
	(0.006)	(0.005)	(0.002)	(0.018)
Outright homeowner dummy	0.001	-0.002	-0.046	-0.030
	(0.009)	(0.008)	(0.032)	(0.026)
Mortgagor dummy	0.010	0.008	0.005	0.007
	(0.009)	(0.007)	(0.032)	(0.025)
Age	0.000	0.000	-0.000	-0.000
	(0.002)	(0.002)	(0.002)	(0.002)
Age-squared	0.000	0.000	-0.000	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Married dummy	0.002	-0.001	0.059**	0.044**
	(0.006)	(0.005)	(0.023)	(0.018)
Number of children	0.001	0.004*	0.001	0.000
	(0.003)	(0.002)	(0.011)	(0.008)
Log liquid wealth	-0.000		0.04	
	(0.001)		(0.004)	
R^2	0.03	0.04	0.02	0.02
Number of observations	739	1,099	720	1,075

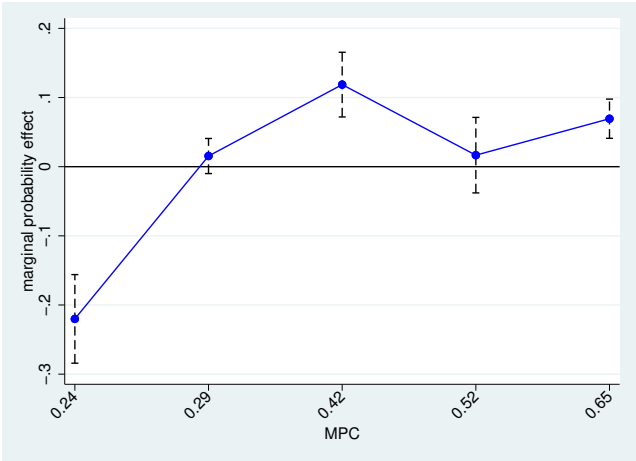
Note: The dependent variable in column (III) and (IV) is a dummy that takes 1 if the modal durable MPC is above 0.5. All logged variables takes 0 when the raw value is 0 or negative. *, ** and *** denote significance of the coefficients at 10, 5 and 1% respectively. Standard errors in parentheses.

Table 11: Multinomial logit on modal MPC out of total expenditures

MPC	0.289	0.422	0.516	0.651
Log salary income	0.011	0.0003	-0.035	0.199
	(0.067)	(0.039)	(0.035)	(0.132)
Log total income	0.976**	1.057***	0.802***	2.530***
	(0.391)	(0.203)	(0.186)	(0.574)
Mortgage interest to income ratio	-0.448	1.399*	2.183***	-2.091
	(2.283)	(0.791)	(0.738)	(2.614)
APC	0.947**	1.316***	0.577***	3.672***
	(0.422)	(0.218)	(0.220)	(0.539)
Outright homeowner dummy	-0.404	-0.000	0.228	0.500
	(0.589)	(0.311)	(0.305)	(0.827)
Mortgagor dummy	-0.408	0.177	0.663**	1.033
	(0.555)	(0.321)	(0.309)	(0.857)
Age	0.012	-0.057	-0.023	-0.009
	(0.089)	(0.042)	(0.041)	(0.110)
Age-squared	0.000	0.001	0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Married dummy	1.179**	0.364*	0.306	0.463
	(0.465)	(0.212)	(0.200)	(0.536)
Number of children	0.025	0.053	0.021	-0.048
	(0.200)	(0.093)	(0.090)	(0.201)
Pseudo R^2		0.08		
Number of observations		1,079		

Notes: Output of a single multinomial logit estimation. The excluded base outcome is the lowest MPC, 0.245. All logged variables takes 0 when the raw value is 0 or negative. *, ** and *** denote significance of the coefficients at 10, 5 and 1% respectively. Standard errors in parentheses.

Figure 16: Marginal and average propensities to consume: multinomial logit



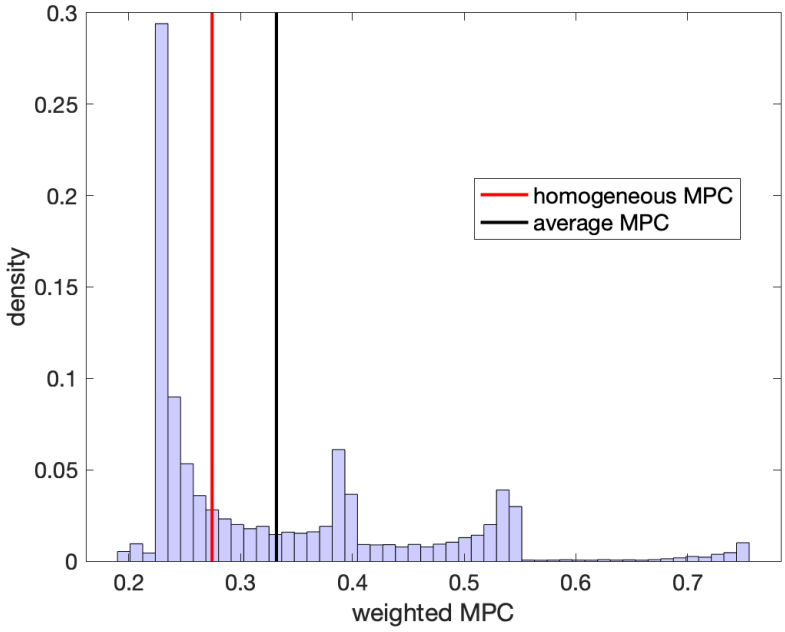
Notes: Figure 16 shows the marginal probability effect of a household's APC, measured as mean lagged consumption relative to lagged total income, on the modal MPC. Total expenditures.

By absorbing the lagged consumption response, this modification accounts for the fact that, in the baseline specification, the control group includes households that received the rebate in the past, and whose consumption response might be persistent. [42] We then interact the rebate, its lagged value, and the constant with the group indicators $\mathbf{1}[j \in g]$, and solve the FCM algorithm to get the endogenous weights and the vector of coefficients $\left\{\theta_g, \theta_g^{\text{lag}}, \alpha_g\right\}$.In Figure 17, we show that the distribution of weighted MPCs is very similar to the one estimated in the baseline specification.

To address the fact that some households never receive the rebate (and thus may be meaningfully different from those who do receive the rebate), we drop households who do not receive a rebate within the sample period we cover.40% of the observations in the sample are associated with households that do not receive a rebate in this time period. Figure 18 shows the distribution of the weighted MPCs in this subsample. Our results are very similar to those shown in Figure 2, with a slight rightward shift of the distribution. Indeed, Parker et al. (2013) also estimate a larger homogeneous rebate coefficient in this subsample.

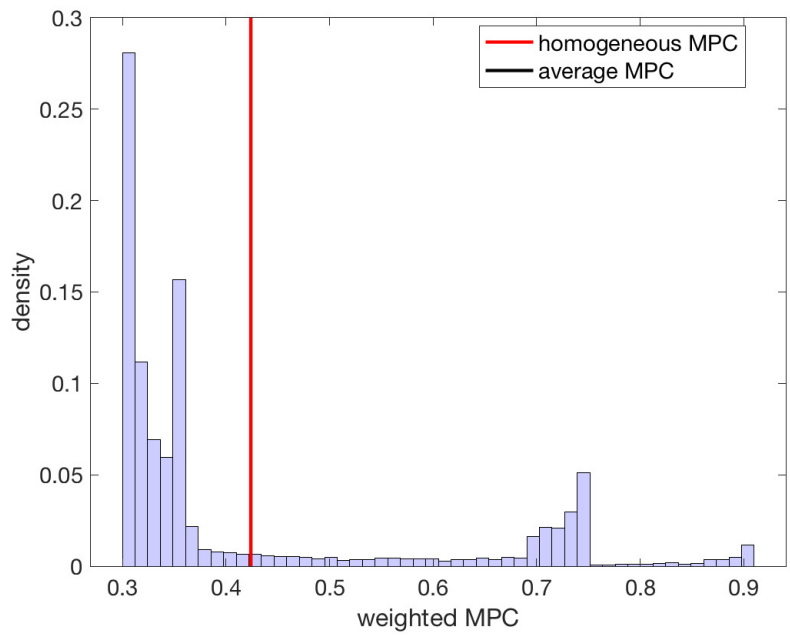
[42]This is true so long as the persistent effect of the rebate lasts strictly less than four quarters. Moreover, we assume that the policy is fully anticipated by all households. In an intermediate information case in which, for instance, the policy enters the agents' information set after the receipt of the first rebate, this specification cannot fully account for anticipatory effects often labelled as MPC out of news.

Figure 17: Estimated distribution of MPCs out of the tax rebate: control for lagged responses



Notes: Total expenditures, defined as in Parker et al. (2013). The homogeneous MPC (red line) is estimated assuming homogeneous response to the tax rebate, as in Parker et al. (2013). For each household we compute the weighted MPC. The black line shows the average weighted MPC in our sample.

Figure 18: Estimated distribution of MPCs out of the tax rebate: only rebate recipients



Notes: Total expenditures, defined as in Parker et al. (2013). The homogeneous MPC (red line) is estimated assuming homogeneous response to the tax rebate, as in Parker et al. (2013). For each household we compute the weighted MPC. The black line shows the average weighted MPC in our sample. Red and black lines overlap almost exactly.

D.4 Homogeneous and average MPC

In a model with heterogeneous effects, it is not generally true that an estimated homogeneous effect is equal to the weighted average of the heterogeneous effects. In our context, this point is important: even if a researcher is interested only in the average MPC, that MPC will not generally be recovered by estimating a homogeneous effect. To see this, consider a simple two-group model of the form

$$y_i = \mathbf{1}\left[i \in D_1\right] \theta_{D_1} x_i + \mathbf{1}\left[i \in D_2\right] \theta_{D_2} x_i + e_i, \tag{22}$$

and assume that observations $i = 1, \ldots, N/2$ are in group D_1 and the remainder are in group D_2. Then, the population counterpart of the standard OLS estimator for a homogeneous slope, $\bar{\theta}$, is given by

$$\bar{\theta} = E\left[x_i^2\right]^{-1} E\left[x_i y_i\right] = \left(E\left[x_i^2 \mid i \in D_1\right] + E\left[x_i^2 \mid i \in D_2\right]\right)^{-1} \left(E\left[x_i^2 \mid i \in D_1\right] \theta_{D_1} + E\left[x_i^2 \mid i \in D_2\right] \theta_{D_2}\right)$$

In this simple model, the true average MPC is $\left(\theta_{D_1} + \theta_{D_2}\right)/2$. However, $\bar{\theta}$ is equal to this value in general if and only if the distribution of x_i is independent of group membership.

In this case, $E\left[x_i^2 \mid i \in D_1\right] = E\left[x_i^2 \mid i \in D_2\right] = E\left[x_i^2\right]$, so

$$\bar{\theta} = \left(2E\left[x_i^2\right]\right)^{-1}\left(E\left[x_i^2\right]\theta_{D_1} + E\left[x_i^2\right]\theta_{D_2}\right) = \frac{\theta_{D_1} + \theta_{D_2}}{2}.$$

However, in general, if researchers hope to recover a causal effect, group membership (and thus the treatment effect θ_{g_i}) should be independent of x_i. In our setting, this means that rebate value should not be correlated with the MPC. Under these assumptions, the homogeneous MPC estimated from (22) will recover the average MPC. However, this is not the case if we consider a more complicated model, with additional regressors. In particular, consider

$$y_i = \mathbf{1}\left[i \in D_1\right]\theta_{D_1}x_i + \mathbf{1}\left[i \in D_2\right]\theta_{D_2}x_i + W_i'\beta + e_i, \tag{23}$$

where W_i is a vector of controls. In this case, the homogeneous coefficients will be given by

$$\begin{aligned}\bar{\theta}_{aug} &= E\left[X_iX_i'\right]^{-1}E\left[X_iy_i\right] \\ &= \frac{E\left[X_iX_i' \mid i \in D_1\right]\theta_{D_2} + E\left[X_iX_i' \mid i \in D_2\right]\theta_{D_2}}{E\left[X_iX_i' \mid i \in D_1\right] + E\left[X_iX_i' \mid i \in D_2\right]}\end{aligned} \tag{24}$$

where $X_i = \begin{pmatrix} x_i \\ W_i \end{pmatrix}$ and $\bar{\theta}_{aug}$ stacks homogeneous coefficients on X_i. Now, for $\bar{\theta}$, the coefficient on x_i, to recover the average effect, the distribution of X_i must be independent of group membership, or W_i must be independent of x_i. This is a much stronger assumption. Indeed, in our setting, this would require that any included controls have no predictive power for an individual's MPC. More broadly, this violates the basis for the entire literature studying MPC heterogeneity correlated with observables. The same argument also holds for the case with additional group-specific coefficients and is easily extended to $G > 2$ groups.