Fast Estimation of Dynamic Structural Models with unobserved heterogeneity*

– preliminary version –

Jeppe Druedahl†
Thomas H. Jørgensen‡
Dennis Kristensen§

September 30, 2020

Abstract

We propose a novel approximate fixed effects (AFE) estimator that employs interpolation in the computation of its criterion function. This feature greatly reduces the number of times the underlying economic model needs to be solved. In the case of dynamic programming models this can reduce the estimation time from days to minutes. We study the asymptotic behavior of the AFE estimator and derive the leading additional biases due to approximations under mild regularity conditions. We demonstrate that the Jackknife removes both the usual incidental parameter bias and biases due to approximations. Monte Carlo results highlights the attractive features of the AFE which is much faster than the exact FE estimator and with only small additional estimation errors. We apply the AFE to fit the buffer-stock consumption-saving model with unrestricted heterogeneity in the discount factor on Danish register data.

*We thank Bo E. Honoré, Elena Manresa, Christopher Carroll, Mette Ejrnæs, Lutz Hendricks, Rasmus Søndergaard Pedersen, Søren Leth-Petersen, Claus Thustrup Kreiner and Anders Munk-Nielsen for fruitful discussions and suggestions. Financial support from the Danish Council for Independent Research in Social Sciences is gratefully acknowledged (FSE, grant no. 4091-00040 and 5052-00086B).

†CEBI, Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 35, DK-1353 Copenhagen K, Denmark. E-mail: jeppe.druedahl@econ.ku.dk. Website: http://econ.ku.dk/druedahl.

‡CEBI, Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 35, DK-1353 Copenhagen K, Denmark. E-mail: thomas.h.jorgensen@econ.ku.dk. Webpage: www.tjeconomics.com.

§Department of Economics, University College London, Gower Street, London, United Kingdom. E-mail: d.kristensen@ucl.ac.uk. Website: https://sites.google.com/site/econkristensen.

Keywords: Heterogeneity, fixed effects, structural estimation, interpolation, consumption-saving.

1 Introduction

Economic agents are typically heterogeneous in terms of ex ante characteristics. Experiments, for example, repeatedly provide evidence of substantial heterogeneity in preferences, abilities and beliefs.[1] Such heterogeneity furthermore have important positive and normative implications. In terms of economic modelling and estimation, it is therefore pivotal to allow for flexible forms of heterogeneity.

In a parametric setting unobserved heterogeneity can be modelled by treating (some of) the parameters as random effects (RE's) or fixed effects (FE's) that vary across individuals. However, most applied structural work assume ex ante homogeneity or impose strong parametric restrictions on the variation, e.g., restrict heterogeneity to be from discrete distriction; see, e.g., French and Jones (2011); Ciliberto and Tamer (2009); Bonhomme and Manresa (2015) These strong parametric restrictions on heterogeneity comes with a significant risk of misspecification with biases in results and conclusions as consequence. When panel data is available, it is well-known that this risk can be removed by treating the individual-specific parameter values as FE's to be estimated together with any common parameters.

But very few empirical studies impement structural models with flexible RE's or FE's. One of the main reasons for this is computation time; it is often not computationally feasible to allow such features. Many structural models cannot be solved on closed form and so their estimation normally involve an outer and inner loop where in the outer one we search over the parameter space and in the inner loop, for a given candidate value of the parameter, numerical dynamic programming is used to solve the model. If the support of the random coefficients or fixed effects is large, the model has to be solved many times and so estimation time becomes prohibitively large.

We propose a general numerical algorithm that resolves this computational issue. As an example, our proposal allows us to estimate the canonical buffer-stock consumption model with discount rates treated as FE's in a matter of minutes for a sample of more than 250,000 households observed over 8 time periods. In comparison, the standard FE estimator takes hours to deliver similar estimates. The basic idea of our approximate estimator is to reduce the time spent in the inner loop of the estimation procedure. This is achieved by precomputing the solution to the economic model on a grid spanning the relevant domain of the heterogeneous parameters and the model's state space. In the subsequent estimation procedure, the model can then be evaluated by interpolation

[1] Examples include Barsky, Juster, Kimball and Shapiro (1997), Coller and Williams (1999), Beetsma and Schotman (2001), Holt and Laury (2005), Andersen, Harrison, Lau and Rutström (2008), Guiso and Paiella (2008), Dohmen, Falk, Huffman, Sunde, Schupp and Wagner (2011), Andreoni and Sprenger (2012) and Finke and Huston (2013).

instead of computing a new solution to the model, which computationally is orders of magnitude more expensive due to use of e.g. dynamic programming or the need to find fixed points. Moreover, partial derivatives of the objective function w.r.t. the variables that interpolation is employed on are available on closed form and can be computed very fast. This means that derivative-based numerical optimizers can be applied without relying on numerical derivatives. Our algorithm proves to be particularly powerful when estimating FE models but we expect it to also be useful in estimation of RE models and, more generally, models with a large number of homogeneous parameters and/or observations.

We provide a general asymptotic theory for approximate estimation in FE and RE panel data models. We apply the general theory to our interpolation method which allows us to derive the additional biases due to interpolation being used in the estimation. We furthermore demonstrate that these biases (together with biases due to FE's) can be removed by Jackknife. The theory is general enough that it can also be used to analyze the effects of other numerical tools, such as simulation-based methods, on estimation of FE and RE models.

We investigate the performance of our method in practice through a set of Monte Carlo experiments. These show that only a modest number of grid points is needed in order for the AFE to be close to identical to the exact estimator. We additionally suggest a simple data driven approach to choose the bounds and density of the grid, where the model solutions are pre-computed.

To illustrate the empirical applicability of our proposed estimator, we estimate the buffer-stock consumption model on Danish administrative register data allowing for heterogeneous discount factors. This model was first structurally estimated in Gourinchas and Parker (2002) and Cagetti (2003) assuming homogeneous preferences. We are the first to estimate the model without making any distributional assumptions on the form of heterogeneity. Our results suggest that there is substantial preference heterogeneity. The importance of allowing for preference heterogeneity to explain wealth inequality is noted by De Nardi and Fella (2017), while Krueger, Mitman and Perri (2016), Carroll, Slacalek, Tokuoka and White (2017) and Alan, Browning and Ejrnæs (2017) study its importance for consumption dynamics. After discussing the related literature below, the paper proceeds as follows. Sections 2 and 5 present the approximate FE (AFE) and approximate RE (ARE) estimators, while Section A contains the asymptotic theory. Section 4 presents the Monte Carlo estimation results. In Section 7, we report the estimation results from our empirical application. Finally, we conclude in Section 8. All proofs and lemmas have been relegated to Appendix A.

1.1 Existing Literature

Interpolation has been used in elsewhere in structural empirical work to obtain a smoothed model solution from a discretized version; see, e.g., Keane and Wolpin (1994) and Low, Meghir and Pistaferri (2010). However, they only interpolate over the state variables while we also use interpolation for the unknown parameters leading to substantial computational savings compared to their method when used in estimation.

Similarly, the idea of pre-computing the model solution on a fixed grid of parameters before estimation goes back to at least the histogram method of Kamakura (1991). Bajari, Fox and Ryan (2007) and Fox, Kim, Ryan and Bajari (2011) use this for simple estimation of static discrete choice models with nonparametric RE's. Fox, Kim and Yang (2016) provide formal justification for this approach. Unfortunately, the procedure becomes much more complex for structural dynamic models because it generally requires solving a high-dimensional non-linear optimization problem with all the population weights as parameters to be estimated.[2] Dynamic models with RE's, furthermore, face the initial condition problem where the researcher must specify how the RE distribution depends on the intial values of the state variables (see e.g. Heckman, 1981). The FE version of our approximate estimator does not face these problems. Finally, Han (????) also proposes an estimator that employs interpolation. He focuses on cross-sectional applications without fixed effects.

In a similar vein, Hahn and Moon (2010); Bonhomme and Manresa (2015) and Bester and Hansen (2015) developed grouped fixed effect (GFE) estimators where the FE's are assumed to have finite support. When applied to structural models, this means that the model only has to be estimated at the support points and so appear to come with similar computational advantages. In fact, in the special case of interpolation being done using step functions, our estimator becomes computationally equivalent to their estimator with our grid point corresponding to the placements of the groups. However, while Bonhomme and Manresa (2015) assume that the finite FE model is the data generating mechanism, we treat it as an approximation to an underlying continuous distribution of the FE's and our theory takes into account the biases due to this. Bonhomme, Lamadon and Manresa (2017) consider the extension to the case where unobserved heterogeneity is not necessarily discrete in the population, and the number of groups therefore is required to grow with the sample size. However, if indeed the underlying distribution is continuous, we recommend using a higher-order (smooth) interpolation scheme since this comes with smaller numerical errors as demonstrated in our theory.

[2] The constrained least squares formulation of the estimator can, as shown by Nevo, Turner and Williams (2016), be recovered for continuous choices in a method of moment version where all the moments are restricted to be linear in the population weights.

The above cited papers belong to a large literature on estimation of mixture models. A particularly popular estimator in this class is the non-parametric maximum likelihood estimator (NPMLE) proposed by Heckman and Singer (1984), among others. These types of estimators often formulate an *expected* likelihood function where both the groups placement and weights are to be estimated. In terms of computation, these estimators can be numerically unstable due to the simultaneous estimation of weights and nodes which can result in multiple local optima and problems of convergence. Empirical applications have therefore been restricted to cases with a few (e.g. 3) distinct groups.

From a methodological perspective, there is a large literature that analyzes the effect of approximations on estimators based on cross-sectional or time series data; see, e.g., Fernández-Villaverde, Rubio-Ramírez and Santos (2006), Kristensen and Salanié (2017) and Kristensen and Shin (2012). To our knowledge, this is the first paper that provide a theory for approximate estimators in a panel data setting. Our asymptotic results extend the ones for FE estimators found in Hahn and Newey (2004) and Hahn and Kuersteiner (2011) to take into account numerical approximations in the computation of the objective functions.

2 Estimating FE models using Interpolation

We here first show how our proposal works when applied to a consumption-saving model in Section 2.1 and then present the general version of our AFE estimator in Section 2.2 and the proposed Jackknife in Section 2.3. Section 2.4 discusses the practical implementation of the AFE.

2.1 Illustrative example: Consumption-saving model

To illustrate how our AFE estimator works, we here explain its implementation in the context of a consumption-saving model where the discount rate is treated as a FE. Specifically, we consider the canonical buffer-stock model of Deaton (1991, 1992) and Carroll (1992, 1997, 2012) where individual i chooses consumption C_{it} to mamize expected utility subject to financial constraints,

$$\begin{aligned} V_t(M_{i,t}, P_{i,t}) &= \frac{C_{i,t}^{1-\rho}}{1-\rho} + \beta_i \mathbb{E}_t[V_{t+1}(M_{i,t+1}, P_{it+1})] \qquad (2.1) \\ &\text{s.t.} \\ A_{it} &= M_{it} - C_{it},\ M_{it+1} = rA_{it} + Y_{it+1},\ P_{it+1} = gP_{it}\psi_{it+1}, \\ Y_{it+1} &= \begin{cases} 0 & \text{with. prob. } \pi \\ P_{it+1}\xi_{it+1} & \text{else} \end{cases}, \end{aligned}$$

and $A_{it} \geq -\lambda P_{it+1}$. Here, the state variables are cash-on-hand M_{it} and permanent income P_{it}, while the shocks satisfy $\log \xi_{it+1} \sim \mathcal{N}(-0.5\sigma_\xi^2, \sigma_\xi^2)$ and $\log \psi_{it+1} \sim \mathcal{N}(-0.5\sigma_\psi^2, \sigma_\psi^2)$. We follow Gourinchas and Parker (2002) and model consumption in the retirement period T_R as

$$C^*_{T_R}(M_{iT}, P_{iT}) = \gamma \frac{1 - R^{-1}(\beta R)^{1/\rho}}{1 - [R^{-1}(\beta R)^{1/\rho}]^{L-T}} \left[M_{iT} + \frac{1 - R^{-(L-T)}}{1 - R^{-1}} \kappa P_{iT} \right] \tag{2.2}$$

where κ is the replacement rate and L is the last period of life. The parameter γ shifts retirement utility and so affects the propensity to consumer in retirement relative to during worklife.

To keep exposition simple, we here fix most of the parameters in the model and only treat the discount rate β_i and the consumption preference γ as free parameters to be estimated. We here assume that γ is common to all inviduals while β_i is treated as a FE. The model can be sold in terms of $c_{i,t} = C_{i,t}/P_{i,t}$ and $m_{i,t} = M_{i,t}/P_{i,t}$ so that $c_{i,t} = c^\star_t(m_{i,t}; \gamma, \beta_i)$. We allow for relative consumption to be observed with error so that

$$c_{i,t} = c^\star_t(m_{i,t}; \gamma, \beta_i) + \varepsilon_{i,t}, \;\; \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_c^2),$$

where $\varepsilon_{i,t}$ is the measurement error with variance σ_c^2, and β_i is individual i's discount parameter.

The consumption function $c^\star_t$ is not available on closed form but a very good numerical approximation can be obtained by the endogeneous grid method (EGM) (Carroll, 2006). EGM takes as input the model parameters (γ, β_i) and returns the values of $c^\star_t(m; \gamma, \beta_i)$ for m on a set of grid points chosen by us. Thus, interpolation is needed anyway in order to compute $c^\star_t(m_{i,t}; \gamma, \beta_i)$ when the observed value $m_{i,t}$ falls between the grid points used in EGM. This was implemented using 500 grid points for the end-of-period asset grid and Gauss-Hermite quadrature with 5 nodes for the income shocks. We will here ignore any numerical errors contained in this approximate solution; but note that our theory accommodates this feature.

Given observations of consumption and savings for a random sample of N individuals over T time periods, the exact FE estimator of the common parameters (γ, σ_c^2) and the FE's $\beta_1, ..., \beta_N$ then solves

$$\begin{aligned} \hat{\gamma} &= \arg\min_{\gamma} \sum_{i=1}^{N} \sum_{t=1}^{T} (c_{it} - c^\star_t(m_{it}; \gamma, \hat{\beta}_i(\gamma)))^2 \\ \hat{\beta}_i(\gamma) &= \arg\min_{\beta} \sum_{t=1}^{T} (c_{it} - c^\star(m_{it}; \gamma, \beta))^2, \; i = 1, \ldots, N, \end{aligned}$$

and $\hat{\sigma}_c^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (c_{it} - c^\star_t(m_{it}; \hat{\gamma}, \hat{\beta}_i(\hat{\gamma})))^2$. The computation of these estimators is costly since, for each value of γ, we have to compute $\hat{\beta}_1(\gamma), ..., \hat{\beta}_N(\gamma)$, and each of these require us to recompute $c^\star(m_{it}; \gamma, \beta)$ again and again as we search over β. In practice,

with $N = 1000$ and $T = 10$, it took us around 100 seconds to compute $\hat{\gamma}$. This may not sound like a lot, but the computational burden quickly increases as more parameters are treated as FE's and/or as N gets bigger.

Our proposal circumvents the computational bottleneck caused by the repeated numerical evaluation of $c_t^\star(m;\gamma,\beta)$: We choose J grid points for each of the two variables m and β, say, $\{m_1,...,m_J\}$ and $\{\beta_1,...,\beta_J\}$, respectively. For a given value of γ, we then compute $c_{j,k}(\gamma) = c_t^\star(m_j;\gamma,\beta_k)$, $j,k = 1,....,J$, and use, e.g., tensor B-splines, to compute the function off the grid,

$$\hat{c}_{K,t}^\star(m;\gamma,\beta) = \sum_{j,k=1}^{K} c_{j,k}(\gamma) B'_{j,k} \left[\sum_{j,k=1}^{K} B_{j,k}B'_{j,k}\right]^{-1} B(m) \otimes B(\beta), \tag{2.3}$$

where $B_{j,k} = B(m_j) \otimes B(\beta_k)$ is the tensor B-spline evaluated at the (j,k)th node. Here, $K = J^2$ is the total number of grid points and so controls the interpolation error and the computation time – a larger value of K reduces the interpolation error but at the same time increases the computation time since it involves K evaluations of $c_t^\star$.

We then replace the consumption function appearing in the FE estimation problem with $\hat{c}_{K,t}^\star$ to obtain our AFE,

$$\begin{aligned}
\hat{\gamma}_K &= \arg\min_{\gamma} \sum_{i=1}^{N}\sum_{t=1}^{T}(c_{it} - \hat{c}_{K,t}^\star(m_{it};\gamma,\hat{\beta}_i(\gamma)))^2 \\
\hat{\beta}_{K,i}(\gamma) &= \arg\min_{\beta} \sum_{t=1}^{T}(c_{it} - \hat{c}_{K,t}^\star(m_{it};\gamma,\beta))^2,\ i = 1,\ldots,N.
\end{aligned}$$

The central benefit of AFE is that, for a given value of γ, the computation of $\hat{\beta}_{K,i}(\gamma)$ involves an objective function which is avaliable on closed form since $\hat{c}_{K,t}^\star(m;\gamma,\beta)$ in eq. (2.3) is on closed form (after pre-computing $c_{j,k}(\gamma)$). When estimating the heterogeneous parameters, the AFE thus interpolates a pre-computed interpolant, which typically is much faster than solving the underlying economic model, as is done in the FE estimator. Note in particular that the number of evaluations of $c_t^\star$ is independent of N.

To demonstrate the computational gains of the AFE in this context, report results from a Monte Carlo experiment. The data-generating parameters was chosen as $\gamma = 1$ while $\beta_i = \min\left\{\max\left\{\tilde{\beta}_i, \underline{\beta}\right\}, \overline{\beta}\right\}$ with $\tilde{\beta}_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$. The remaining parameters of the model are, as mentioned earlier, treated as known; the values of these can be found in Table 1. We then simulate $N = 1000$ individuals who are observed from age 40–49 ($T = 10$); all individuals are born with no wealth ($A_{i0} = 0$) and with permanent income normalized to one ($P_{i0} = 1$).

For the interpolation, we used three different basis functions, cubic splines, linear splines and step functions to investigate the sensitivity to the type of interpolation. For all

Table 1: Monte Carlo experiment: Data-generating parameter values

ρ	γ	r	g	κ	λ	σ_ψ	σ_ξ	π	$\underline{\beta}$	$\overline{\beta}$	μ_β	σ_β^2	σ_c^2	T	L
2.0	1.0	1.03	1.02	0.9	0.0	0.1	0.1	0.01	0.90	0.99	0.95	0.01	0.1	40	60

three, we used uniform grids with the bounds of the m-grid chosen as the 1st and 99th percentile of the data, while the bounds for the α-grid were 0.80 and 1.05.[3] We furthermore investigate how J affects the performance by varying this between 5 and 100.

The results for the homogeneous parameter is shown in Table 2 and Figure 4. Table 2 reports respectively the bias, the standard deviation across Monte Carlo runs, the root mean squared error, the average estimation time and the number of times $c_t^\star$ had to be computed for various choices of the interpolant and number of grid points, J. The rows of Table 1 labelled $\hat{J}$ shows Monte Carlo results when choosing J using a data-driven algorithm described in Section 2.4.3 below, where we also discuss the corresponding numerical.

We see that, as J increases, our AFE estimator converges to the FE estimator irrespective of the choice of the interpolation scheme. In terms of estimation time, we see that the bi-linear interpolation approach is the fastest for a given J. The cubic spline, however, converges faster implying that the estimation time for the lowest J where convergence to the FE estimator is ensured are very similar across these two interpolation approaches. Classification is slower. It should, however, be noted that the computational cost of the cubic spline for high J increases more than linearly in J, while bi-linear interpolation increases less than linearly due to the approximate fixed costs of searching for the optimal heterogeneous parameters given the bi-linear interpolant.[4]

Comparing computation times, we see that, with J chosen so that the MSE of AFE is comparable to the MSE of the exact one, our AFE estimators are roughly 16 times faster than the FE estimator when using splines ($J = 50$), 40 times faster for linear interpolation ($J = 100$), and 15 times faster when using classification ($J = 250$). This result become even more stark as N increases, as we will see in our application to Danish register data in Section 7. The last column shows that our AFE estimator requires much fewer solutions of the dynamic programming problem. If the model was harder to solve, the speed-up of our AFE estimator relative to FE would consequently increase.

The results for the heterogeneous parameter is shown in Table 6 and Figure 5. Table 6 reports the average root mean squared error and its standard deviation across Monte Carlo runs for various J. We again see the same convergence patterns as for the homogeneous

[3] These bounds were chosen to ensure that no households were estimated with discount factors outside these bounds.

[4] An additional downside of the cubic spline is that the implementation is more complicated.

parameter. Figure **??** shows the distributions of heterogeneous parameter pooled across Monte Carlo runs. We see that the distribution of the heterogeneous parameters for our AFE estimators converges to that of the FE estimator as J increases, which itself is almost correctly centered, but have excessive dispersion.

Table 2: Example 2. Homogeneous, γ.

	Bias	MC std.	RMSE	Time (secs)	Solutions
FE	0.007	0.028	0.029	99.1	91829
AFE, Cubic spline interpolation					
$J = 5$	-0.399	0.132	0.420	2.8	87
$J = 10$	0.024	0.025	0.034	1.7	97
$J = 25$	0.009	0.029	0.030	2.3	213
$J = 50$	0.007	0.028	0.029	5.7	418
$J = 100$	0.007	0.028	0.029	8.3	832
$\hat{J} = 43.8$ (avg.)	0.007	0.029	0.029	9.7	580
AFE, Linear interpolation					
$J = 5$	-0.605	0.049	0.607	3.9	120
$J = 10$	-0.111	0.028	0.115	1.4	80
$J = 25$	-0.022	0.026	0.034	1.4	181
$J = 50$	0.000	0.028	0.028	1.7	398
$J = 100$	0.005	0.028	0.029	2.5	820
$J = 250$	0.006	0.028	0.029	4.8	2084
$\hat{J} = 87.8$ (avg.)	0.004	0.028	0.029	9.0	1552
AFE, Classification					
$J = 5$	0.798	0.047	0.800	5.9	208
$J = 10$	0.319	0.252	0.406	3.4	234
$J = 25$	0.047	0.045	0.065	2.2	264
$J = 50$	0.015	0.031	0.034	2.1	450
$J = 100$	0.008	0.029	0.030	3.5	892
$J = 250$	0.007	0.029	0.029	6.5	2153
$J = 500$	0.006	0.029	0.029	12.4	4248
$\hat{J} = 94.5$ (avg.)	0.008	0.029	0.030	11.0	1781

Notes: Shows Monte Carlo results for Example 2 for the homogeneous parameter, γ with $N = 1000, T = 10$. We have used 250 Monte Carlo runs and the parameters in Table 1.

Table 3: Example 2. Heterogeneous, β_i.

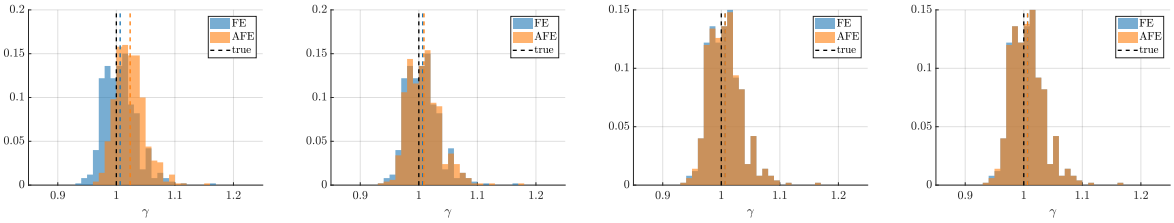
	Avg. RMSE	MC std.
FE	1.284	0.035
AFE, Cubic spline interpolation		
$J = 5$	11.475	0.075
$J = 10$	1.352	0.034
$J = 25$	1.279	0.034
$J = 50$	1.284	0.035
$J = 100$	1.284	0.035
$\hat{J} = 43.8$ (avg.)	1.284	0.035
AFE, Linear interpolation		
$J = 5$	14.795	0.029
$J = 10$	3.353	0.060
$J = 25$	1.407	0.039
$J = 50$	1.303	0.036
$J = 100$	1.288	0.035
$J = 250$	1.284	0.035
$\hat{J} = 87.8$ (avg.)	1.290	0.035
AFE, Classification		
$J = 5$	2.604	0.029
$J = 10$	1.527	0.038
$J = 25$	1.324	0.036
$J = 50$	1.293	0.035
$J = 100$	1.286	0.035
$J = 250$	1.284	0.035
$J = 500$	1.284	0.035
$\hat{J} = 94.5$ (avg.)	1.286	0.035

Notes: Shows Monte Carlo results for Example 2 for the heterogeneous parameter, β_i with $N = 1000, T = 10$. We have used 250 Monte Carlo runs and the parameters in Table 1.

Figure 1: Example 2. Homogeneous, γ.

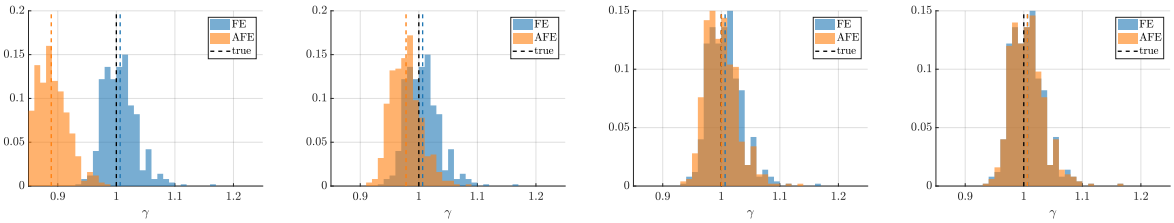
Cubic spline interpolation

(a) $J = 10$ (b) $J = 25$ (c) $J = 50$ (d) $J = 100$



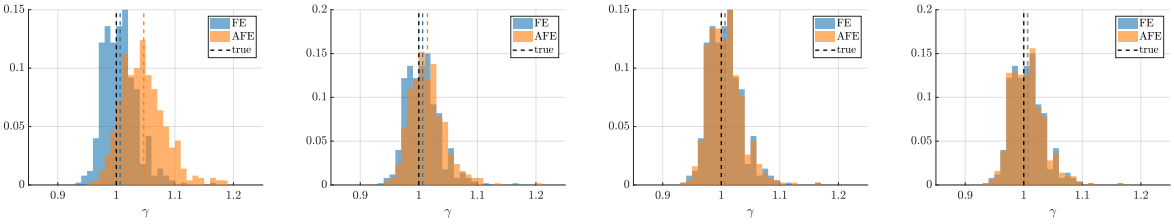
Linear interpolation

(e) $J = 10$ (f) $J = 25$ (g) $J = 50$ (h) $J = 250$



Classification

(i) $J = 25$ (j) $J = 50$ (k) $J = 250$ (l) $J = 500$

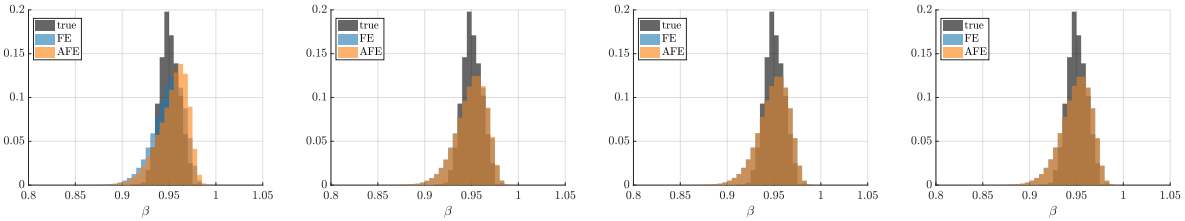


Notes: Shows Monte Carlo results for Example 2 for the homogeneous parameter, γ for selected J with $N = 1000, T = 10$.. The dashed black line shows the true value. The remaining dashed lines show the means of, respectively, the FE and AFE estimators. We have used 250 Monte Carlo runs and the parameters in Table 1.

Figure 2: Example 2. Heterogeneous, γ.

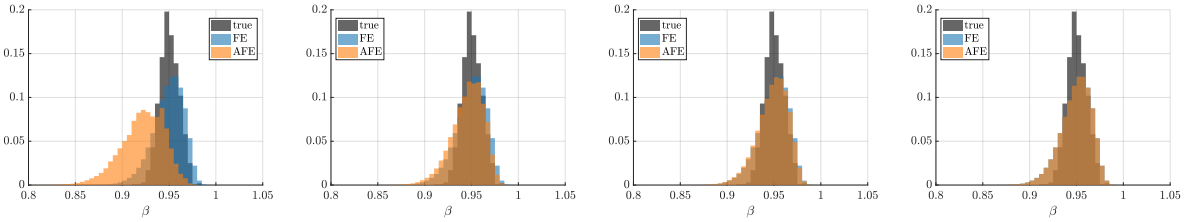
Cubic spline interpolation

(a) $J = 10$ (b) $J = 25$ (c) $J = 50$ (d) $J = 100$



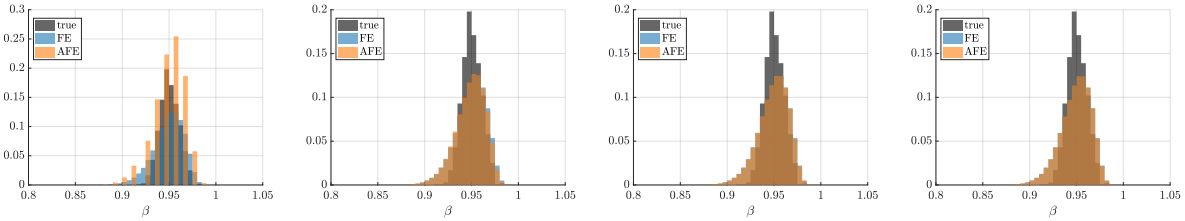
Linear interpolation

(e) $J = 10$ (f) $J = 25$ (g) $J = 50$ (h) $J = 250$



Classification

(i) $J = 25$ (j) $J = 50$ (k) $J = 250$ (l) $J = 500$



Notes: Shows Monte Carlo results for Example 2 for the heterogeneous parameter, β_i for selected J with $N = 1000, T = 10$.. The distributions of the heterogeneous parameter, β_i, are pooled across Monte Carlo runs. We have used 250 Monte Carlo runs and the parameters in Table 1.

2.2 General framework

We now present our proposal in a general setting where we take as given a structural model characterized by a "solution" $\psi\left(x_{i,t};\theta_1,\alpha_{1,i}\right)$ where θ_1 is a set of common parameters and $\alpha_{1,i}$ contains the FE's. In addition to $x_{i,t}$ the researcher also observe a set of output variables $y_{i,t}$ which we collect in $z_{i,t} = (y_{i,t}, x_{i,t})$, $i = 1, ..., N$ and $t = 1, ..., T$.

Given model and data, the researcher has developed an objective function

$$q(z_{i,t};\theta,\alpha_i,\psi) := r(y_{i,t},\psi\left(x_{i,t};\theta_1,\alpha_{1,i}\right);\theta_2,\alpha_{2,i}), \tag{2.4}$$

for some function r and $\psi_{i,t}\left(\theta_1,\alpha_{1,i}\right) = \psi\left(x_{i,t};\theta_1,\alpha_{1,i}\right)$. The solution may depend on ad-

ditional model parameters and FE's as captured by $(\theta_2, \alpha_{2,i})$. We collect the common parameters in $\theta = (\theta_1, \theta_2)$ and the FE's in $\alpha_i = (\alpha_{1,i}, \alpha_{2,i})$, $i = 1, ..., N$. The objective function is application specific; it could, for example, be a non-linear least-squares estimator in which case $r(y_{i,t}, \psi_{i,t}; \theta_2, \alpha_{2,i}) = (y_{i,t} - \psi_{i,t})^2/\sigma^2$, where $\theta_2 = \sigma^2$ is the error variance. But many other types of estimators are allowed for, including maximum-likelihood.

Given q, the *"exact" fixed effects estimator (FE)* is defined as

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta\in\Theta} \sum_{i=1}^{N}\sum_{t=1}^{T} q(z_{it}; \theta, \alpha_i(\theta), \psi) \\ \hat{\alpha}_i(\theta) &= \arg\min_{\alpha\in\mathcal{A}} \sum_{t=1}^{T} q(z_{it}; \theta, \alpha, \psi), \ \forall i = 1, \ldots, N, \end{aligned} \tag{2.5}$$

where by exact we mean that $\psi\left(x_{i,t}; \theta_1, \alpha_{1,i}\right)$ is computed without error. But if the computation of ψ is costly, our *approximate fixed effects estimator (AFE)* may be an attractive alternative; this takes the form

$$\begin{aligned} \hat{\theta}_K &= \arg\min_{\theta\in\Theta} \sum_{i=1}^{N}\sum_{t=1}^{T} q(z_{it}; \theta, \alpha_{J,i}(\theta), \hat{\psi}_K) \\ \alpha_{K,i}(\theta) &= \arg\min_{\alpha\in\mathcal{A}} \sum_{t=1}^{T} q(z_{it}; \theta, \alpha, \hat{\psi}_K), \ \forall i = 1, \ldots, N, \end{aligned} \tag{2.6}$$

where $\psi\left(x_{i,t}; \theta_1, \alpha_{1,i}\right)$ is replaced by its interpolant,

$$\hat{\psi}_K(x; \theta_1, \alpha_1) = \sum_{j,k=1}^{J} \psi(x_j; \theta, \alpha_{1,k}, \psi) B'_{j,k} \left[\sum_{j,k=1}^{K} B_{j,k} B'_{j,k}\right]^{-1} B\left(x\right) \otimes B\left(\alpha\right), \tag{2.7}$$

based on $K = J^2$ grid points. Here, for simplicity, we assume the same number of grid points J are used for each of the variables. In practice, one may wish to use different grid points depending on the curvature of the solution w.r.t. the different variables and the "size" of the space that is being interpolated over; see Section 2.4.3 for further details.

Instead of first interpolating ψ and the plugging this back into the objective function, one could employ interpolation on $q(z; \theta, \alpha, \psi)$ itself. If, for example, we again interpolate w.r.t. x and α_i, this would take the form

$$\hat{q}_K(z; \theta, \alpha, \psi) = \sum_{j,k=1}^{J} q(y, x_j; \theta, \alpha_{1,k}, \alpha_2, \psi) B'_{j,k} \left[\sum_{j,k=1}^{J} B_{j,k} B'_{j,k}\right]^{-1} B\left(x\right) \otimes B\left(\alpha\right). \tag{2.8}$$

This has the same computational cost as $\hat{\psi}_J(x; \theta_1, \alpha_1)$ but is linear in the interpolation error which has certain advantages in terms of the bias it induces in $\hat{\alpha}_{K,i}$, $i = 1, ..., N$, and $\hat{\theta}_K$. In Section 3, we show that the interpolant in (2.8) generally will suffer from fewer biases compared to the one in (2.7) because the approximation appears linearly in

the objective function in the former case.

So far we have focused on objective functions on the form (2.4). However, our AFE estimator applies more generally to objective functions $q(y_{it};\theta,\alpha,\psi)$ that depend implicitly on some underlying function ψ which is costly to compute. For example, it could take the form $q(z_{i,t},\theta,\alpha,\psi)=q(y_{i,t},m_{i,t}\left(\theta_2,\alpha_{i,2},\psi\right);\theta_1,\alpha_{i,1})$, where $m_{i,t}=m\left(x_{1,i,t};\theta_{21},\alpha_{i,21},\psi\left(x_{2,i,t};\theta_{22},\alpha_{i,22}\right)\right)$. If the function m is numerically cheap and dim $\left(x_{2,i,t};\theta_{22},\alpha_{i,22}\right)$ is smaller than dim $\left(x_{1,i,t};\theta_{21},\alpha_{i,21},\psi\right)$,it may be computationally advantageous to interpolate ψ instead of m (or q). However, again, this solution will tend to generate additional biases in the corresponding objective function if m exhibits strong non-linearities in ψ. This more general version also allows for, e.g., latent dynamic variables that have to be integrated out in the computation of the objective function, and other cases where the computationally expensive component of the model enters the objective function in a more complex manner.

Observe that the interpolation leads to reduced computation time of the objective function defining the estimators. It also also for simple computation its partial derivatives w.r.t. θ and α_i,....,α_N. For the "direct" interpolator in eq. (2.8), its first-order derivative w.r.t. α is given by

$$\frac{\partial\hat{q}_K(z;\theta,\alpha)}{\partial\alpha}=\sum_{j,k=1}^{J}q(z_j;\theta,\alpha_k,\psi)B'_{j,k}\left[\sum_{j,k=1}^{K}B_{j,k}B'_{j,k}\right]^{-1}B\left(x\right)\otimes\frac{\partial B\left(\alpha\right)}{\partial\alpha}.$$

This means that derivative-based optimizers can be used to compute the AFE with high precision.

2.3 Jackknife

As can be seen from the results for the consumption model, the AFE will generally suffer from additional biases due to interpolation. We here propose to use the so-called Jackknife to remove some of these biases. The basic ("half-panel") Jackknife splits the panel in two subsamples along the time dimension, $\{z_{i,t}:t=1,....,T/2,i=1,....,n\}$ and $\{z_{i,t}:t=T/2,....,T,i=1,....,n\}$, where we for simplicity assume T is even, and then re-estimate θ and the FE's based on each of the two subsamples. With $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ denoting the two additional estimates, the Jackknife estimator then takes the form

$$\tilde{\theta}=2\hat{\theta}-\frac{1}{2}\left(\hat{\theta}^{(1)}+\hat{\theta}^{(2)}\right). \tag{2.9}$$

In order for the Jaccknife to work, it is important that $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are computed using the exact same interpolation scheme, incl. number and location of gridpoints. This also means that the computation of the two additional estimators come at a low additional computationl cost.

The Jackknife has already been shown to remove the so-called incidental parameter bias that most FE estimators suffer from; see Hahn and Newey (2004) and Dhaene and Jochmans (2015). In Section 3, we extend their theory and show that the Jackknife at the same time also moves any interpolation biases in the AFE.

Similar to Hahn and Newey (2004), one can alternatively attempt to estimate the leading bias term due to approximation. This requires knowledge of the pointwise interpolation error (instead of just a bound for it). This is unfortunately not available in general. One exception of this is when B-splines are being used in which case the interpolation error is known, c.f. Zhou and Wolfe (2000). Due to its computational simplicity, we here advocate using Jackknife.

NUMERICAL RESULTS....

2.4 Implementation in practice

2.4.1 Which variables to include in interpolation

The researcher is free to decide which of the variables in $(x_{i,t}; \theta_1, \alpha_{1,i})$ that interpolation is employed. To speed up computation, one may wish to interpolate all variables. However, this comes at a cost of precision: Keeping the total number of grid points J fixed, the more variables that are included the bigger the interpolation error will become; see Section 3 for further details. Thus, the choice depends on how the researcher values computation time over numerical precision.

2.4.2 Choice of interpolation method

The numerical literature offers a wide range of interpolation schemes, but standard choices are B-splines, Legendre polynomials, Hermite polynomials and Chebyshev polynomials. We refer to Judd (1998) for an introduction to these. We found that B-splines were particularly useful, however, since they and their partial derivatives are fast to compute and come with added degrees of freedom in terms of smoothness. As shown in the Monte Carlo study, one can use cubic B-splines which are twice differentiable; or first-order B-splines (linear interpolation) which are continuous but non-differentiable; or "zero-order" B-splines which correspond to step functions. The last category corresponds to classification as discussed in the introduction.

2.4.3 Choosing the Grid

In order to implement our AFE estimator it is necessary for the econometrician to chose a grid, which we denote $\mathcal{G}_K$, on which we pre-compute the solution to the model. We propose to make this choice by a simple data driven approach. We first choose the bounds

of the grid heuristically, and then choose the number of grid points in each dimension by a simple algorithm. For simplicity, we restrict attention to equally spaced tensor product grids with the same number of grid points, J, in each dimension, though nothing in the AFE estimator requires this.[5]

In choosing the bounds of the grid, we can firstly use that in the data dimensions (the z dimension), we ex ante know where we will need to evaluate the interpolant. For the heterogeneous parameters (the α dimension) we instead propose to choose the grid bounds to ensure that no or very few units are estimated to be on the boundary of the grid. For preference parameters in particular, the econometrician typically has valuable prior information about the domain, but otherwise a trial-and-error approach can be used.[6]

Given the grid bounds and the assumption of equally spaced tensor product grids with J grid points in each dimension, we can write the criterion function for a given guess θ as a function of J,

$$\hat{Q}_{N,J}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} q(z_{it}, \hat{\alpha}_{J,i}(\theta); \theta, \hat{\psi}_J), \tag{2.10}$$

where we index the interpolator $\hat{\psi}_J$ as a function of J. We propose to determine J using that our AFE estimator converge to the FE estimator as J is increased (see Section A). This in particular implies that the change in the estimated objective function should go towards zero when increasing J. Hence, we propose to choose J as the smallest J where the objective function does not change any more.

To be specific, define the maximum change in the objective function over an l step window, with a step size of Δ, as

$$\delta(J, l, \Delta) = \max_{k \in \{1,2,\dots,l\}} \left| \frac{\hat{Q}_{N,J-(k-1)\Delta}(\theta) - \hat{Q}_{N,J-k\Delta}(\theta)}{\Delta} \right|$$

We then propose to determine J as the smallest J where this maximum change is below some tolerance

$$\hat{J} = \arg\min \left\{ J : \ \delta(J, l, \Delta) < \eta, \ J = k\Delta, \ k \in \{l+1, l+2, \dots\} \right\}, \tag{2.11}$$

The termination tolerance, η, is similar to termination tolerances employed when performing numerical optimization in general. We suggest setting η to either 10^{-4} or 10^{-5}.

Note that all of the above is done for a single guess of θ. If the non-linearity of the model

[5] It is, for example, possible to use non-tensor non-equally spaced grids such as adaptive sparse grids. This might in particular be interesting in high-dimensional settings.

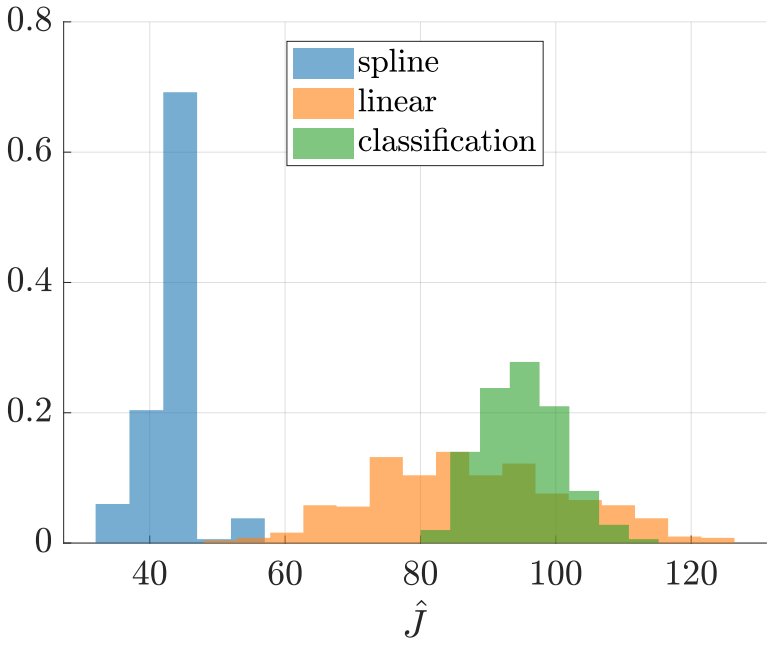
[6] If there are local minima with respect to α, choosing too narrow bounds could result in the estimator returning a local rather than global minimum even if there is no observations on the boundary of the fixed parameter space.

vary drastically with θ it might be beneficial to try different guesses and pick the highest $\hat{J}$. Alternatively, it can be checked that the $\hat{J}$ implied by the estimated $\hat{\theta}$ is in line with the $\hat{J}$ chosen based on the initial guess.[7]

The performance of the method for the consumption model can be found in Table 1 in the rows labelled $\hat{J}$. With the cubic spline this procedure results in no difference between the FE and AFE estimator at the third decimal place, while it for linear interpolation and the classification approach results in very small deviations. At convergence, and including the time to determine $\hat{J}$, our AFE estimators are roughly 10 times faster than the FE estimator when using spline or linear interpolation, and 8 times faster when using classification.

Figure 3 shows the distribution of $\hat{J}$ based on our Monte Carlo study of the consumption model. For the cubic spline the choice is almost always 45, while for linear interpolation and classification it fluctuates around 60–120 and 80–115, respectively.

Figure 3: Example 2. Histogram of $\hat{J}$ by interpolation method.



Notes: Shows Monte Carlo results for Example 2 for estimated J with $N = 1000, T = 10$. We have used 250 Monte Carlo runs and the parameters in Table 1. The length of the moving average is $l = 3$, the step size $\Delta = 5$ and the tolerance $\eta = 10^{-5}$.

[7] Finally, one could replace θ with the estimated value $\hat{\theta}_J$ in (2.10). This would lead to re-estimation of all model parameters for each guess of J.

3 Asymptotic Theory

We here develop a general asymptotic theory for fixed effects (FE) estimators where an functional component of the model of interest is either approximated or estimated. As a special case, the theory covers the proposed interpolation method. We first introduce some notation: For any given N-dimensional vector $a = (a_1, ..., a_N)$, we write $\|a\|_\infty = \max_{i=1,...,N} \|a_i\|$. For any given $(N \times T)$-matrix $(a_{i,t})$, $i = 1, ..., N$ and $t = 1, ..., T$, we write $\bar{a}_i = \sum_{t=1}^T a_{i,t}/T$. For any given function $a(z_{i,t}, \theta, \alpha_i, \psi)$, we write $a_{0,i,t}(\psi) = a(z_{i,t}, \theta_0, \alpha_i(\theta_0), \psi)$, $\bar{a}_i(\theta, \psi) = \sum_{t=1}^T a(z_{i,t}, \theta, \alpha_i(\theta), \psi)/T$, $\bar{a}_{0,i}(\psi) = \bar{a}_i(\theta_0, \psi)$, $A_i(\theta, \psi) = E[\bar{a}_i(\theta, \psi)]$, $A_{0,i}(\psi) = A_i(\theta_0, \psi)$, and $A_0(\psi) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N A_{0,i}(\psi)$.

3.1 Framework

We first consider a class of standard FE estimators without any approximations. We are given a criterion function $q_{i,t}(\theta, \alpha, \psi) = q(z_{i,t}; \theta, \alpha, \psi)$ that identifies the parameters of interest in the sense that

$$\theta_0 = \arg\min_{\theta\in\Theta} Q(\theta, \psi_0), \tag{3.1}$$

where ψ_0 is the "true" value of some underlying component and, using the notation introduced earlier,

$$Q(\theta, \psi_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N Q_i(\theta, \alpha_{0,i}(\theta, \psi_0), \psi_0),$$

with $Q_i(\theta, \alpha, \psi) = E[q_{i,t}(\theta, \alpha, \psi)]$, $i = 1, ..., N$, and

$$\alpha_{0,i}(\theta, \psi_0) = \arg\min_{\alpha\in\mathcal{A}} Q_i(\theta, \alpha, \psi_0). \tag{3.2}$$

Suppose that the true value of the nuisance parameter, ψ_0, is unknown but we are given an approximation of it, $\hat{\psi}$. This may be the interpolator described in the previous section but we allow for other types of approximations due to, e.g., simulation, discretization, etc. Then the following fixed-effects (FE) extremum estimator is the natural sample analogue to the above population quantities,

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \frac{1}{N} \sum_{i=1}^N \bar{q}_i\left(\theta, \hat{\alpha}_i\left(\theta, \hat{\psi}\right), \hat{\psi}\right), \tag{3.3}$$

where, for $i = 1, ..., N$,

$$\hat{\alpha}_i\left(\theta, \hat{\psi}\right) = \arg\min_{\alpha_i\in\mathcal{A}} \bar{q}_i\left(\theta, \alpha, \hat{\psi}\right), \ \bar{q}_i(\theta, \alpha, \psi) := \frac{1}{T} \sum_{t=1}^T q_{i,t}(\theta, \alpha, \psi). \tag{3.4}$$

Note here that, for notational convenience, we neither index the approximator $\hat{\psi}$ nor the AFE's by J, where now J should be thought of more generally as the degree of approximation being used in the computation of $\hat{\psi}$.

We extend the asymptotic theory of Hahn and Newey (2004) and Hahn and Kuersteiner (2011) [HK, henceforth] to take into account the presence of $\hat{\psi}$; this could be a function, a finite-dimensional parameter of fixed dimension, or a set of pre-estimated fixed effects in which case its dimension grows with N. We assume that $\hat{\psi}$ and ψ_0 are situated in a normed space $(\Psi, \|\cdot\|)$ and will then require that $q_{i,t}(\theta, \alpha, \psi)$ and relevant derivatives of this function are Lispchitz w.r.t. ψ:

Definition 3.1. We say that a given function $f(z; \theta, \alpha, \psi)$ is L_p-Lipshitz continuous w.r.t. ψ for $p \geq 1$ if it satisfies $\|f(z; \theta, \alpha, \psi) - f(z; \theta, \alpha, \psi_0)\| \leq B_f(z) \|\psi - \psi_0\|$ for some function $B_f(z)$ with $\max_{i=1,...,N} E\left[B_f^p(z_{it})\right] < \infty$ and for all ψ in a (small) neighbourhood of ψ_0.

Assumption 1. *(i)* $\{z_{i,t} : t = 1, 2, ...T, i = 1, ..., N\}$ *satisfy Condition 3 in HK; (ii)* $q_{i,t}(\theta, \alpha, \psi_0)$ *satisfies Conditions 1 and 4-7 in HK; (iii) for all* $\psi \in \Psi$ *in a small neighborhood of* ψ_0*,* q *and its partial derivatives up to order 3 are all* L_p-Lipshitz continuity w.r.t. ψ with $p > 10(1 + p_0)/(1 - 10v)$, where $p_0 \geq (d_\theta + d_\alpha + 4)/2$.

We refer to HK for a detailed discussion of the conditions imposed there. The main requirements are that, for each $i = 1, ..., N$, $\{z_{i,t} : t = 1, 2, ...\}$is stationary and geometrically α-mixing; that $(\theta, \alpha) \mapsto q_{i,t}(\theta, \alpha, \psi_0)$ and its partial derivatives satisfy suitable moments; and that the it identifies θ_0 and $\alpha_i(\theta_0)$, $i = 1, ..., N$, as $N, T \to \infty$, c.f. eqs. (3.1)-(3.2). Importantly, HK's conditions imply a set of limit results as found in HK; for convenience, these can be found in Lemma A.2

Note that Assumption 1(ii)-(iiii) implicitly imposes smoothness conditions on $\hat{\psi}$ and ψ_0: If these are functions of (α, θ), then they must necessarily be three times continuously differentiable w.r.t. (α, θ) in order for Assumption 1 to hold. Next, we require the approximator to converge sufficiently fast:

Assumption 2. *(i)* $\left\|\hat{\psi} - \psi_0\right\| = o_P(1)$*; (ii)* $\left\|\hat{\psi} - \psi_0\right\| = O_P(T^{-\rho})$ *with* $\rho := 4/10 - v$ *for some* $0 < v < 1/10$.

The assumption is a high-level one in order to allow for a wide range of approximation schemes. It will have to be verified for the particular scheme being used; see Section X for examples of this. The assumption allows the approximator to be potentially random, which is the case if simulation-based approximation methods are employed. This combined with the Lipschitz condition in Assumption 1 allow us to generalize some of the limit results in HK to allow for the presence of $\hat{\psi}$; see Lemmas A.1. Lemmas A.1 and A.2 will be used in the following to establish the asymptotic properties of the above class of approximate estimators.

We first show consistency of the approximate estimators and develop a higher-order expansion of the FE estimators:

Theorem 3.1. *Under Assumptons 1-2, the following hold, $i = 1, ..., N$:*

$$\hat{\alpha}_{0,i}\left(\hat{\psi}\right) = \alpha_{0,i}\left(\psi_0\right) + \bar{u}_i\left(\hat{\psi}\right) + \bar{v}_i\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right) + r_i^{(\alpha)}, \tag{3.5}$$

where $\bar{u}\left(\hat{\psi}\right)$ and $\bar{v}\left(\hat{\psi}\right)$ are defined in eqs. (A.8)-(A.9). They satisfy $\left\|\bar{u}\left(\hat{\psi}\right)\right\|_\infty = O_P\left(T^{-\rho}\right)$, $\left\|\bar{v}\left(\hat{\psi}\right)\right\|_\infty = O_P\left(T^{-\rho}\right)$ and $\left\|r^{(\alpha)}\right\|_\infty = O_P\left(T^{-3\rho}\right)$ with $\rho > 0$ defined in Assumption 2.

As can be seen from the rate results, $\bar{u}_i\left(\hat{\psi}\right)$ and $\bar{v}_i\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right)$ are the first- and second-order terms corresponding to the leading variance and bias term of $\hat{\alpha}_{0,i}\left(\hat{\psi}\right)$. This higher-order expansion generalizes the one derived in HK to allow for the presence of a first-step estimator/approximator $\hat{\psi}$; see also p. 1303 in Hahn and Newey (2004). The leading terms are identical to the ones in HK, except that they now are functionals of $\hat{\psi}$.

We then use this expansion of the FE estimators to obtain one for $\hat{\theta}$. As a first step, we follow HK and expand $\hat{\theta}$ w.r.t. $\hat{\alpha}_i\left(\theta_0\right)$; see Theorem A.2. Next, we wish toexpand $\hat{\theta}$ w.r.t. $\hat{\psi}$. To do so, we need the relevant components of the expansion in Theorem A.2 to be smooth functionals of ψ. Formally, we introduce the following concept:

Definition 3.2. A function$f(z;\theta,\alpha,\psi)$ is twice L_p-differentiable, $p \geq 1$, w.r.t. ψ at ψ_0 if there exits functionals $\nabla f(z;\theta,\alpha)[d\psi]$ and $\nabla^2 f(z;\theta,\alpha)[d\psi,d\psi]$ which are linear and bi-linear w.r.t. $d\psi$, respectively, and both L_p-Lipschitz w.r.t. $d\psi$ so that for any ψ in a neighbourhood of ψ_0, with $d\psi = \psi - \psi_0$,

$$\left\|f(z;\theta,\alpha,\psi) - f(z;\theta,\alpha,\psi_0) - \nabla f(z;\theta,\alpha)[d\psi] - \frac{1}{2}\nabla^2 f(z;\theta,\alpha)[d\psi,d\psi]\right\| \leq B_f(z)\left\|d\psi\right\|^2$$

for some function $B_f(z)$ with $\max_{i=1,...,N} E\left[B_f^p(z_{it})\right] < \infty$.

Similar conditions can be found in the asymptotic theory for semi- and nonparametric estimators; see, e.g., Chen (2007). We then impose the following additional assumption where $s_{0,i,t}^{(\theta)}(\psi)$ and $s_{0,i,t}^{(\alpha)}(\psi)$ denote the partial derivative of q w.r.t. θ and α_i, respectively.

Assumption 3. *The functions $s_{0,i,t}^{(\theta)}(\psi)$ and $s_{0,i,t}^{(\alpha)}(\psi)$* are twice L_p-differentiable w.r.t. ψ.

This allows us to obtain our first main result:

Theorem 3.2. *Under Assumptions 1-3,*

$$\hat{\theta} - \theta_0 = \Phi_N\left(\psi_0\right) + B_N\left(\psi_0\right)/T + \nabla\Phi_N\left[\hat{\psi} - \psi_0\right] + \frac{1}{2}\nabla^2\Phi_N\left[\hat{\psi} - \psi_0, \hat{\psi} - \psi_0\right] + O_P\left(T^{-3\rho}\right), \tag{3.6}$$

where the four leading terms are defined in eqs. (A.11), (A.12), (A.17) and (A.18), respectively. These satisfy:

$$\sqrt{NT}\Phi_N(\psi_0) \to^D N\left(0, H_0^{(\theta,\theta)}(\psi_0)^{-1}\Omega(\psi_0)H_0^{(\theta,\theta)}(\psi_0)^{-1}\right), \tag{3.7}$$

$$B_N(\psi_0) \to^P B(\psi_0), \tag{3.8}$$

$$\nabla\Phi_N\left[\hat{\psi}-\psi_0\right] = O_P\left(T^{-\rho}\right), \nabla^2\Phi_N\left[\hat{\psi}-\psi_0, \hat{\psi}-\psi_0\right] = O_P\left(T^{-2\rho}\right), \tag{3.9}$$

where the limits can be found in (A.20) and (A.21).

The two terms $\Phi_N(\psi_0)$ and $B_N(\psi_0)/T$, are identical to the leading variance and bias terms derived in HK for the exact FE estimator ($\hat{\psi} = \psi_0$). The two additional terms, $\nabla\Phi_N\left[\hat{\psi}-\psi_0\right]$ and $\nabla^2\Phi_N\left[\hat{\psi}-\psi_0, \hat{\psi}-\psi_0\right]$, contain the leading bias and variance terms due to approximation. The behavior of these depend on the particular type of approximation being used and how the approximated term enter q.

If interpolation is employed, $\nabla\Phi_N\left[\hat{\psi}-\psi_0\right]$ and $\nabla^2\Phi_N\left[\hat{\psi}-\psi_0, \hat{\psi}-\psi_0\right]$ will contain the first and second order interpolation biases; the second order term can therefore be ignored if only the leading bias component is of interest. If simulations are employed so that $E\left[\hat{\psi}\right] = \psi_0$, $\nabla\Phi_N\left[\hat{\psi}-\psi_0\right]$ will contain the leading variance term while $\nabla^2\Phi_N\left[\hat{\psi}-\psi_0, \hat{\psi}-\psi_0\right]$ will contain the leading bias term. If ψ enters the score function linearly, $\nabla^2\Phi_N[d\psi] = 0$ and no second-order effect will be present.

3.2 Jackknife Correction

We here analyze the Jackknife of the AFE that we proposed in eq. (2.9). Hahn and Newey (2004) and Dhaene and Jochmans (2015) showed that the Jackknife for the exact FE estimator removes the leading bias term due to the fixed effects, $B_N(\psi_0)/T$. We here extend their result and show that it will in fact also remove the first-order effect due to approximations. The reason for this is quite intuitive: We can write the adjustment term for $\hat{\theta}$ as $\nabla\Phi_N[d\psi] = \sum_{t=1}^T b_{N,t}[d\psi]/T$ where

$$b_{N,t}[d\psi] = H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{1}{N}\sum_{i=1}^N\left\{\nabla s_{0,i,t}^{(\theta)}[d\psi] + H_{0,i}^{(\theta,\alpha)}(\psi_0)\nabla u_{i,t}[d\psi]\right\},$$

c.f. eq. (A.17). Importantly, if the Jackknife estimator is implemented with *the same* $\hat{\psi}$ being used to compute $\hat{\theta}$, $\hat{\theta}_1$ and $\hat{\theta}_2$, we have that the leading adjustment terms for $\hat{\theta}_1$ and $\hat{\theta}_2$ are given by $\nabla\Phi_N^{(1)}\left[\hat{\psi}-\psi_0\right] = 2\sum_{t=1}^{T/2} b_{N,t}[d\psi]/T$ and $\nabla\Phi_N^{(2)}\left[\hat{\psi}-\psi_0\right] = 2\sum_{t=T/2}^{T} b_{N,t}[d\psi]/T$, respectively. Importantly, we have the following identity,

$$2\nabla\Phi_N[d\psi] - \frac{1}{2}\left(\nabla\Phi_N^{(1)}\left[\hat{\psi}-\psi_0\right] + \nabla\Phi_N^{(2)}\left[\hat{\psi}-\psi_0\right]\right) = 0.$$

Thus, $\tilde{\theta}$ will not contain any first-order term due to the approximation:

Corollary 3.1. *Suppose that Assumptions 1 and 3 hold and* $\left\|\hat{\psi}-\psi_0\right\|^2 = o_P\left(1/\sqrt{NT}\right)$. *Then the Jackknife applied to the approximate FE estimator yields:*

$$\sqrt{NT}\left\{\tilde{\theta}-\theta_0\right\} \to^D N\left(0, H_0^{(\theta,\theta)}\left(\psi_0\right)^{-1}\Omega\left(\psi_0\right)H_0^{(\theta,\theta)}\left(\psi_0\right)^{-1}\right). \tag{3.10}$$

The added rate requirement imposed on $\hat{\psi}$ in the corollary ensures that the second-order term $\nabla^2\Phi_N$ is negiglible. If $\nabla^2\Phi_N = 0$, the requirement can be dropped.

3.3 Grid selection

TBC

3.4 Applications

We here apply the general theory to our interpolation-based AFE. We first consider the case where we "directly" interpolate q and then the "indirect" version. We also demonstrate how our theory is more generally applicable by applying it to simulation-based FE estimators.

3.4.1 "Direct" AFE

We here analyze the AFE when interpolation is employed directly on q leading to $\hat{q}_K$ in (2.8). In this case, $\psi(z_{i,t};\theta,\alpha) := q(z_{i,t};\theta,\alpha,\psi)$ so that $\hat{\psi}(z_{i,t};\theta,\alpha) = \hat{q}_K(z_{i,t};\theta,\alpha)$ is the interpolated objective function based on a total of J grid points. In order to apply our general theory, we need to objective function and its interpolated version to be sufficiently smooth in (θ,α). Formally, we will assume that q and its approximation belong to a so-called Hölder space. Let $f : \mathcal{X} \mapsto \mathbb{R}$, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, be $\underline{\beta} \geq 0$ times differentiable. For any vector $b = (b_1, ..., b_{d_Z}) \in \mathbb{N}_0^{d_X}$ with $|b| = b_1 + \cdots + b_{d_z} \leq \underline{\beta}$, let $D^b f(x) = \partial^{|b|} f(x) / \left(\partial x_1^{b_1} \cdots \partial x_{d_x}^{b_{d_x}}\right)$ be the corresponding partial derivative. For a given $\underline{\beta} \leq \beta < \underline{\beta}+1$, we then define

$$\|f\|_{\beta,\infty} = \max_{|b|\leq\underline{\beta}} \left\|D^b f\right\|_\infty + \max_{|b|=\underline{\beta}} \sup_{x_1\neq x_2} \frac{\left|D^b f(x_1) - D^b f(x_2)\right|}{\|x_1 - x_2\|^{\beta-\underline{\beta}}}, \tag{3.11}$$

where $\left\|D^b f\right\|_\infty = \sup_{x\in\mathcal{X}}\left|D^b f(x)\right|$, and let $\mathbb{C}_{\beta,r}(\mathcal{X})$ be the Hölder space containing all $\underline{\beta} \geq 0$ times continuously differentiable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{\beta,\infty} < r$, where $r \leq +\infty$. To allow for a wide range of interpolation schemes, we then impose the following high-level assumption on the particular interpolation method in use:

Assumption 4. *With* $q_0(z;\theta,\alpha_i) = q(z;\theta,\alpha_i,\psi_0)$ *being the exact objective function, the interpolation error satisfies* $\|\hat{q}_K - q_0\|_{1,\infty} = O(K^{-\gamma_0})$ *and* $\|\hat{q}_K - q_0\|_{3,\infty} = O(K^{-\gamma_1})$*for some* $\gamma_0, \gamma_1 > 0$.

This assumption implicity requires the objective function to be bounded. This is a strong assumption. One can weaken this by using weighted versions of the sup-norm in eq. (3.11), $\left\|D^b f\right\|_\infty = \sup_{x\in\mathcal{X}} \left|w(x) D^b f(x)\right|$. Here, w should then be chosen such that q and its derivatives are bounded in this norm and $E[1/w(z_{i,t})] < \infty$.

The rate with which the interpolation error goes to zero is governed by γ_0 and γ_1; ideally these should be large so that the error vanishes quickly as we increase the total number of interpolation points K. Generally, we have $\gamma_0 > \gamma_1$ since they govern the error rates for the interpolated first-order and the second+third-order partial derivatives, respectively, of q w.r.t. (θ,α). This condition is satisfied for a wide range of finite-dimensional function approximators, including interpolation based on higher-order B-splines (Zhou and Wolfe, 2000), Lagrange polynomials (Howell, 1991) and Hermite polynomials (Birkhoff and Priver, 1967). Suppose that the objective function q and its interpolant $\hat{q}_J$ belong to $\mathbb{C}_{\beta,r}(\mathcal{Z}\times\Theta\times\mathcal{A})$, $\beta \geq 3$ and $r < \infty$, and interpolation is done with either of these choices of bases where we use J polynomial terms in each dimension. Then Assumption 4 holds with $K = J^{d_I}$, where d_I is the number of variables that we interpolate over, $\gamma_0 = (\beta-2)/d_I$ and $\gamma_1 = (\beta-3)/d_I$. Observe that γ_0 and γ_1 increase with the degree of smoothness β; the more smooth q is, the smaller the interpolation error. Reversely, γ_0 and γ_1 decrease as we increase the number of variables that we interpolate over, d_I. Thus, interpolation suffers from a computationaly curse-of-dimensionality: To reach a given level of error tolerance, we need to increase K exponentially with d_I.

Next, we need expressions of the differentials of $\bar{s}_{0,i}^{(\alpha)}(q) = \partial q/(\partial\alpha)$ and $\bar{s}_{0,i}^{(\theta)}(q) = \partial q/(\partial\theta)$ since these enter eqs. (A.17)-(A.19). The first order term is obtained by plugging $\nabla\bar{s}_{0,i}^{(\alpha)}[dq] = d\bar{q}_i^{(\alpha)}$ and $\nabla\bar{s}_{0,i}^{(\theta)}[dq] = d\bar{q}_i^{(\theta)}$, where $dq^{(\alpha)} = \partial(dq)/(\partial\alpha)$ and similar for the other terms, into the expression of $\nabla\Phi_N$, while the second-order term is zero, $\nabla^2\Psi_N = 0$. Thus, there no remainder term in the expansion in eq. (A.16) and so Assumption 3 is trivially satisfied. Moreover, there is no second-order bias term due to interpolation.

Assumption 4 implies Assumption 2 if we restrict $K = K_T \to \infty$ sufficiently fast as $T\to\infty$. That is, we require $K^{\gamma_1} = O(T^\rho)$. We conclude:

Corollary 3.2. *Suppose that Assumptions 1(i)-(ii) and 4 hold where* $K^{-\gamma_1}T^\rho \to 0$ *and* $J^{-\gamma_0} \simeq T^{-1}$. *Then*

$$\hat{\theta} - \theta_0 = \Phi_N(q_0) + \nabla\Phi_N[\hat{q}_K - q_0] + B_N(q_0)/T + O_P\left(T^{-3\rho}\right), \tag{3.12}$$

where, with $\hat{s}^{(\theta)}_{K,i,t}$ *and* $\hat{u}_{K,i,t}$ *denoting the interpolated versions of* $s^{(\theta)}_{0,i,t}$ *and* $u_{i,t}$,

$$\nabla\Phi_N\left[\hat{q}_K - q_0\right] = H_0^{(\theta,\theta)}\left(\psi_0\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\left\{\left(\hat{\bar{s}}^{(\theta)}_{K,i} - \bar{s}^{(\theta)}_i\right) + H_{0,i}^{(\theta,\alpha)}\left(\psi_0\right)\left(\bar{u}_{K,i} - \bar{u}_i\right)\right\} = O_P\left(T^{-1}\right).$$

Furthermore, the Jackknife version in (2.9) satisfies (3.10).

3.4.2 "Indirect" AFE

Consider now instead the case where the ARE is based on interpolation of the solution which is then plugged into the objective function. While our theory applies to a broad class of objective functions, we here focus on the case where q is on the form of (2.4) to avoid complicated notation and assumptions.

First, observe here that the requirement of HK that q is thrice differentiable w.r.t. (θ, α_i) entails that r in (2.4) is thrice differentiable w.r.t. ψ. Similarly, the L_p-Lipschitz condition imposed on q and its partial derivatives w.r.t. (θ, α_i) is satisfied as long as r and its partial derivatives are Lipschitz w.r.t. ψ.

Next, we impose the following high-level assumption on the interpolated solution:

Assumption 5. *The interpolated model solution satisfies* $\left\|\hat{\psi}_K - \psi_0\right\|_{1,\infty} = O\left(K^{-\gamma_0}\right)$ *and* $\left\|\hat{\psi}_K - \psi_0\right\|_{3,\infty} = O\left(K^{-\gamma_1}\right)$ *for some* $\gamma_0, \gamma_1 > 0$.

The discussion of Assumption 4 carries over with obvious modification. In particular, we require for simplicity that the solution and interpolator are bounded functions, but this can be weakened by using weighted norms.

Finally, we need expressions of the differentials of $\bar{s}^{(\alpha)}_{0,i}\left(\psi\right)$ and $\bar{s}^{(\theta)}_{0,i}\left(\psi\right)$ appearing in eqs. (A.17)-(A.19). For notational simplicity, suppose that ψ is a scalar function here. We only present the ones for the FE component since the ones for the common parameters are on a similar form. First note that

$$s^{(\alpha_1)}_{0,i,t}\left(\psi\right) = r^{(\psi)}_{0,i,t}\left(\psi\right)\psi^{(\alpha_1)}_{i,t},\ s^{(\alpha_2)}_{0,i,t}\left(\psi\right) = r^{(\alpha_2)}_{0,i,t}\left(\psi\right),$$

where $r^{(\psi)}_{0,i,t}\left(\psi\right) = \partial r(y_{i,t}, \psi_{i,t}; \theta_{0,2}, \alpha_{2,i})/(\partial\psi)$ and $\psi_{i,t} := \psi_{i,t}\left(\theta_{0,1}, \alpha_{0,1,i}\right)$, and similar for other partial derivatives. The corresponding differentials become

$$\begin{aligned}\nabla s^{(\alpha_1)}_{0,i,t}\left[d\psi\right] &= r^{(\psi)}_{0,i,t}\left(\psi_0\right)d\psi^{(\alpha_1)}_{i,t} + r^{(\psi,\psi)}_{0,i,t}\left(\psi_0\right)\psi^{(\alpha_1)}_{0,i,t}d\psi_{i,t},\\ \nabla s^{(\alpha_2)}_{0,i,t}\left[d\psi\right] &= r^{(\alpha_2,\psi)}_{0,i,t}d\psi_{i,t},\end{aligned} \tag{3.13}$$

and

$$\begin{aligned}\nabla^2 s^{(\alpha_1)}_{0,i,t}\left[d\psi, d\psi\right] &= 2r^{(\psi,\psi)}_{0,i,t}\left(\psi_0\right)d\psi^{(\alpha_1)}_{i,t}d\psi_{i,t} + r^{(\psi,\psi,\psi)}_{0,i,t}\left(\psi_0\right)\psi^{(\alpha_1)}_{0,i,t}\left(d\psi_{i,t}\right)^2,\\ \nabla^2 s^{(\alpha_2)}_{0,i,t}\left[d\psi\right] &= r^{(\alpha_2,\psi,\psi)}_{0,i,t}\left(d\psi_{i,t}\right)^2,\end{aligned} \tag{3.14}$$

We then impose the following regularity conditions in order for Assumption 3 to hold:

Assumption 6. *$r(y, \psi; \theta_1, \alpha_{1,i})$ and its partial derivatives w.r.t. $(\psi, \theta_1, \alpha_{1,i})$ up to order 3 are bounded uniformly in $(\psi; \theta_1, \alpha_{1,i})$ by some function $B(y)$ which has pth moment.*

This condition will hold under great generality as long as the solution mapping $\psi(\cdot)$ is bounded. As in the previous section, we expect that the following result will also hold for unbounded solution mappings; this will however require us to work with weighted norms and involvr more complicated assumptions:

Corollary 3.3. *Suppose that Assumptions Assumptions 1(i)-(ii) and 5-6 hold where $K^{-\gamma_1}T^{\rho} \to 0$ and $J^{-\gamma_0} \simeq T^{-1}$. Then (3.6) holds with the differentials $\nabla\Phi_N$ and $\nabla^2\Phi_N$ given in terms of the differentials in (3.13)-(3.14). Furthermore, the Jackknife version in (2.9) satisfies (3.10).*

Importantly, the indirect interpolator will suffer from additional bias terms due to $\nabla^2\Phi_N \neq 0$ in general unless r is linear w.r.t. ψ, c.f. (3.14).

3.4.3 Simulated ARE with "direct" interpolation

We could in principle here employ the general theory developed in Section 3 to the ARE. However, RE estimators generally do not need $T \to \infty$ in order for a regular asymptotic theory to hold. Specifically, the incidental parameter biases will not be present when heterogeneity is modelled parameterically. We here instead develop a fixed T asymptotic theory for the simulated ARE in (**??**) that takes into account the joint effect of simulations and interpolation.

To simplify notation in our asymptotic analysis, we relabel the components entering the simulated ARE. First, without loss of generality, rewrite the RE's as $\alpha_i = a(u_i; z_{0,i}, \theta_2)$ for some mapping a and where u_i is drawn from a parameter independent distribution $F_u(u)$, for example, the uniform distribution. Here, θ_2 contains any shape parameters of the RE distribution. Next, let

$$f_{Z|\alpha}(Z_i, u_i; \theta) := \prod_{t=1}^{T} f_{z|\alpha}(z_{it}; \theta_1, a(u_i; z_{0,i}, \theta_2), \psi),$$

where $Z_i = (z_{i,0}, ..., z_{i,T})$, $\theta = (\theta_1, \theta_2)$ and we suppress dependence on the solution ψ to the underlying model, denote the conditional likelihood of the ith observational unit, so that the simulated and interpolated version of the unconditional likelihood of Z_i,

$$f_Z(Z_i; \theta, \beta) = \int f_{Z|\alpha}(Z_i, u; \theta)\, dF_u(u),$$

takes the form

$$\hat{f}_Z(Z_i; \theta, \beta) = \frac{1}{S}\sum_{s=1}^{S} \hat{f}_{Z|\alpha}(Z_i, u_s; \theta),$$

where $\hat{f}_{Z|\alpha}$ is an interpolated version of $f_{Z|\alpha}$. Thus, $\hat{f}_Z$ contains both a bias and variance component. Furthermore define

$$q_i^{(1)}(\theta) = \frac{\partial \log f_Z(Z_i;\theta)}{\partial \theta}, \; q_i^{(2)}(\theta) = \frac{\partial^2 \log f_Z(Z_i;\theta)}{\partial \theta \partial \theta'}.$$

We impose the following regularity conditions on the model and simulator:

Assumption 7. *(i) For all ψ in a neighbourhood of ψ_0, $f_{Z|\alpha}(Z_i, u_s; \theta)$ is twice continuously differentiable w.r.t. θ; (ii) $\Theta \times \mathcal{B}$ is compact with (θ_0, β_0) situated in the interior with $E[\log f_Z(Z_i;\theta_0)] > E[\log f_Z(Z_i;\theta)]$ for all $\theta \neq \theta_0$; (iii) $\left|\log f_{Z|\alpha}(Z_i;\theta)\right| \leq B(Z_i)$ where $E[B(Z_i)] < \infty$; (iv) $E\left[q_i^{(1)}(\theta_0)\right] = 0$ and $\Omega = E\left[q_i^{(1)}(\theta_0) q_i^{(1)}(\theta_0)'\right]$ exists; (v) $\left\|q_i^{(1)}(\theta)\right\| \leq B(Z_i)$ for all θ in a neighbourhood of θ_0 and $H_0 = E\left[q_i^{(2)}(\theta_0)\right]$ has full rank; (vii) Z_i has compact support and $f_Z(Z_i;\theta) > 0$ for all (Z_i, θ).*

Parts (i)-(vi) are quite standard for the analysis of standard MLE's. Part (vii) will allow us to control the effects of simulation and interpolation with the main restriction is the compact support assumption; this is used to ensure that the functional derivatives of the log-likelihood w.r.t. $\hat{f}_Z$ are regular. This restriction could be removed if we introduce trimming in the simulated likelihood, c.f. Kristensen and Shin (2012), but this would lead to more complicated arguments and conditions.

To distinguish between the bias and variance component in $\hat{f}_Z$, we introduce the interpolation operator

$$\Pi_K(f)(z,u,\theta) = \sum_{k=1}^K f(z_k,u_k,\theta_k) B_k' \left[\sum_{j,k=1}^K B_k B_k'\right]^{-1} B(z,u,\theta),$$

where $B_k = B(z_k, u_k, \theta_k)$.

Theorem 3.3. *Under Assumption 7,*

$$\hat{\theta} - \theta_0 = \Phi_N + \nabla\Phi_N\left[\hat{\psi} - \psi_0\right] + \frac{1}{2}\nabla^2\Phi_N\left[\hat{\psi} - \psi_0, \hat{\psi} - \psi_0\right] + O_P\left(T^{-3\rho}\right), \tag{3.15}$$

where

$$\begin{aligned}
\sqrt{N}\Phi_N &= H_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N q_i^{(1)}(\theta_0) \to^D N(0,\Omega), \\
\nabla\Phi_N\left[\hat{\psi} - \psi_0\right] &= H_0^{(\theta,\theta)}(\psi_0)^{-1} \frac{1}{N} \sum_{i=1}^N \frac{1}{f_{Z,i}} \left\{E_u\left[q_{i,s}^{(\theta)}(\theta_0)\right] - \right\},
\end{aligned}$$

the four leading terms are defined in eqs. (A.11), (A.12), (A.17) and (A.18), respectively.

These satisfy:

$$\sqrt{N}\Phi_N(\psi_0) \rightarrow^D N\left(0, H_0^{(\theta,\theta)}(\psi_0)^{-1}\Omega(\psi_0)H_0^{(\theta,\theta)}(\psi_0)^{-1}\right), \tag{3.16}$$

$$B_N(\psi_0) \rightarrow^P B(\psi_0), \tag{3.17}$$

$$\nabla\Phi_N\left[\hat{\psi}-\psi_0\right] = O_P\left(T^{-\rho}\right), \nabla^2\Phi_N\left[\hat{\psi}-\psi_0, \hat{\psi}-\psi_0\right] = O_P\left(T^{-2\rho}\right), \tag{3.18}$$

where the limits can be found in (A.20) and (A.21).

4 More Monte Carlo Experiments

In the following Monte Carlo experiments, we examine the performance of AFE when applied to a simple dynamic panel regression model with heterogeneous slopes,

$$y_{it} = \theta\alpha_i^2 + \alpha_i y_{it-1} + \varepsilon_{it}, \ \varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{4.1}$$

The non-standard parametrization of the constant term, $\theta\alpha_i^2$, ensures that the model function is non-linear in the heterogeneous parameter, which allows us to discuss the choice of smoothness of the interpolation approach. We wish to estimate the homogeneous and the heterogeneous parameters using a least squares criteria, $q(z_{i,t};\theta,\alpha_i,\psi) = (y_{it} - \psi(y_{it-1},\hat{\alpha}_i(\theta),\theta))^2$ with $\psi(y_{it-1},\alpha_i,\theta) = \theta\alpha_i^2 + \alpha_i y_{it-1}$.

The chosen parameters are shown in Table 4. In the simulations we use data after a $1,000$ period burn-in, and draw the heterogeneous parameter from a mixture of two truncated normal distributions, $\alpha_i = \min\{\max\{\tilde{\alpha}_i, 0\}, \overline{\alpha}\}$ where $\tilde{\alpha}_i = I\{u_i \leq \pi\}\mathcal{N}(\mu_{\alpha 1}, \sigma_{\alpha 1}^2) + I\{u_i > \pi\}\mathcal{N}(\mu_{\alpha 2}, \sigma_{\alpha 2}^2)$ and $u_i \sim U[0,1]$. We set $N = 1000$ and $T = 10$ and the bounds of the $y_{i,t-1}$-grid are chosen as the 1st and 99th percentile of the data, while the bounds for the α-grid are 0 and 1.

The results for the homogeneous parameter is shown in Table **??** and Figure **??**. Table **??** reports respectively the bias, the standard deviation across Monte Carlo runs, and the root mean squared error for various choices of interpolant and the number of grid points, J. We see that the AFE estimator based on a cubic spline has already converged to the FE estimator with $J = 3$ (because the model function is quadratic in the parameters). With bi-linear interpolation we instead need $J = 50$ for convergence, and with classification we need at least $J = 500$. As expected, convergence is thus faster for more smooth interpolation approaches.

The bias for the AFE estimator is the sum of the well-known incidental parameter bias of the FE estimator and an additional approximation bias, which disappears as J is increased. We see that the approximation bias for the current example is of the opposite sign of the incidental parameter bias. The bias is thus coincidentally lowest for small

J. Figure **??** also shows that the distribution of the AFE estimates of the homogeneous parameter converges to that of the FE estimator as J is increased for all three interpolation approaches. Table 9 and 10 in the Supplemental Material show that for high enough T the half-panel jack-knife of Dhaene and Jochmans (2015) reduce the bias of the FE estimator and the AFE estimator when it has converged to the FE estimator (before convergence the bias can both decrease and increase from applying the half-panel Jackknife).

The results for the heterogeneous parameter are shown in Table **??** and Figure **??**. Table **??** reports respectively the average root mean squared error and its standard deviation across Monte Carlo runs for the various interpolants and J's. We see the same convergence patterns as for the homogeneous parameter. Figure **??** shows the distributions of the heterogeneous parameter pooled across Monte Carlo runs. We see that the distributions of the heterogeneous parameter for our AFE estimators converge to that of the FE estimator, which itself is downward biased due the incidental parameter bias.

Table 4: Example 1. Parameters.

θ	σ_ε	π	$\overline{\alpha}$	$\mu_{\alpha 1}$	$\sigma^2_{\alpha 1}$	$\mu_{\alpha 2}$	$\sigma^2_{\alpha 2}$
0.1	0.02	0.6	0.95	0.65	0.03	0.40	0.03

Table 5: Example 2. Homogeneous, γ.

	Bias	MC std.	RMSE
FE	0.007	0.028	0.029
AFE, Cubic spline interpolation			
$J = 5$	-0.399	0.132	0.420
$J = 10$	0.024	0.025	0.034
$J = 25$	0.009	0.029	0.030
$J = 50$	0.007	0.028	0.029
$J = 100$	0.007	0.028	0.029
$\hat{J} = 43.8$ (avg.)	0.007	0.029	0.029
AFE, Linear interpolation			
$J = 5$	-0.605	0.049	0.607
$J = 10$	-0.111	0.028	0.115
$J = 25$	-0.022	0.026	0.034
$J = 50$	0.000	0.028	0.028
$J = 100$	0.005	0.028	0.029
$J = 250$	0.006	0.028	0.029
$\hat{J} = 87.8$ (avg.)	0.004	0.028	0.029
AFE, Classification			
$J = 5$	0.798	0.047	0.800
$J = 10$	0.319	0.252	0.406
$J = 25$	0.047	0.045	0.065
$J = 50$	0.015	0.031	0.034
$J = 100$	0.008	0.029	0.030
$J = 250$	0.007	0.029	0.029
$J = 500$	0.006	0.029	0.029
$\hat{J} = 94.5$ (avg.)	0.008	0.029	0.030

Notes: Shows Monte Carlo results for Example 2 for the homogeneous parameter, γ with $N = 1000, T = 10$. We have used 250 Monte Carlo runs and the parameters in Table 1.

Table 6: Example 2. Heterogeneous, β_i.

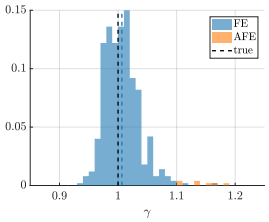
	Avg. RMSE	MC std.
FE	1.284	0.035
AFE, Cubic spline interpolation		
$J = 5$	11.475	0.075
$J = 10$	1.352	0.034
$J = 25$	1.279	0.034
$J = 50$	1.284	0.035
$J = 100$	1.284	0.035
$\hat{J} = 43.8$ (avg.)	1.284	0.035
AFE, Linear interpolation		
$J = 5$	14.795	0.029
$J = 10$	3.353	0.060
$J = 25$	1.407	0.039
$J = 50$	1.303	0.036
$J = 100$	1.288	0.035
$J = 250$	1.284	0.035
$\hat{J} = 87.8$ (avg.)	1.290	0.035
AFE, Classification		
$J = 5$	2.604	0.029
$J = 10$	1.527	0.038
$J = 25$	1.324	0.036
$J = 50$	1.293	0.035
$J = 100$	1.286	0.035
$J = 250$	1.284	0.035
$J = 500$	1.284	0.035
$\hat{J} = 94.5$ (avg.)	1.286	0.035

Notes: Shows Monte Carlo results for Example 2for the heterogeneous parameter, β_i with $N = 1000, T = 10$. We have used 250 Monte Carlo runs and the parameters in Table 1.

Figure 4: Example 2. Homogeneous, γ.

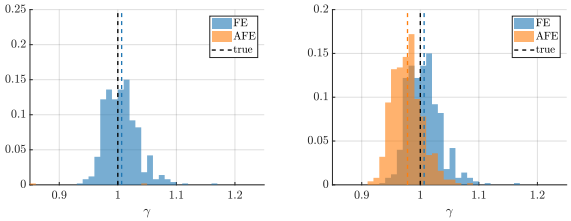
Cubic spline interpolation

(b) $J = 3$ (d) $J = 5$



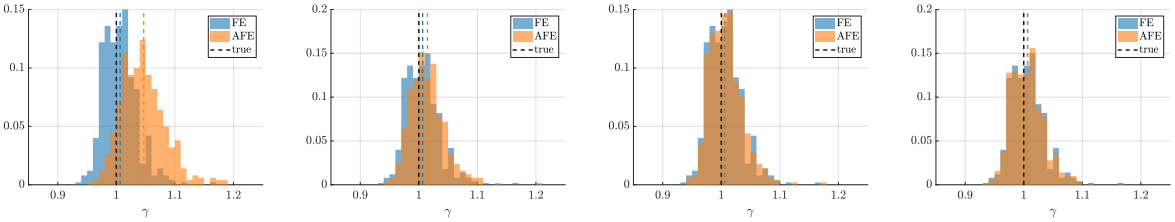
Linear interpolation

(f) $J = 3$ (g) $J = 5$ (h) $J = 25$



Classification

(i) $J = 25$ (j) $J = 50$ (k) $J = 100$ (l) $J = 500$

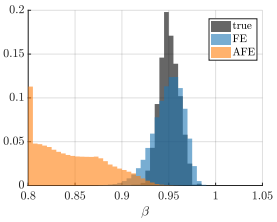


Notes: Shows Monte Carlo results for Example 2 for the homogeneous parameter, γ for selected J with $N = 1000, T = 10$. The dashed black line shows the true value. The remaining dashed lines show the means of, respectively, the FE and AFE estimators. We have used 250 Monte Carlo runs and the parameters in Table 1.

Figure 5: Example 2. Heterogeneous, γ.

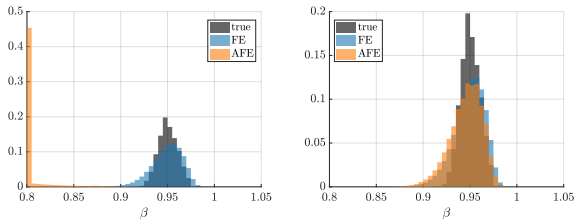
Cubic spline interpolation

(b) $J = 3$ (d) $J = 5$



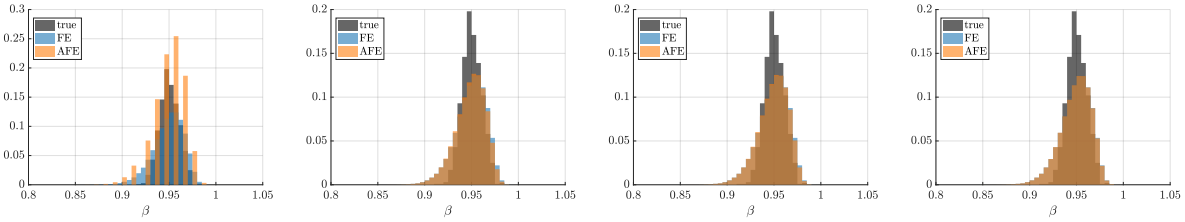
Linear interpolation

(f) $J = 3$ (g) $J = 5$ (h) $J = 25$



Classification

(i) $J = 25$ (j) $J = 50$ (k) $J = 100$ (l) $J = 500$



Notes: Shows Monte Carlo results for Example 2 for the heterogeneous parameter, β_i for selected J with $N = 1000, T = 10$. The distributions of the heterogeneous parameter, β_i, are pooled across Monte Carlo runs. We have used 250 Monte Carlo runs and the parameters in Table 1.

5 Approximate Random Effects (ARE)

We here show how the interpolation method also has uses in estimation of RE models. We here only present the method for the case of simulated maximum-likelihood estimation (SMLE); it should be obvious how to adjust our proposal to handle, e.g., simulated method of moments and other estimation methods for RE models [REFERENCES.....].

We start with a likelihood function of data $f_{z|\alpha}(z_{it}; \theta, \alpha_i, \psi)$, but now treat $\alpha_i \in \mathcal{A}$ as a

RE with known distribution $F_\alpha(\alpha_i|z_{i0};\beta)$, where $\beta \in \mathcal{B}$ are unknown shape parameters to be estimated together with θ. The *"exact" RE estimator* then takes the form

$$\left(\hat{\theta},\hat{\beta}\right) = \arg\min_{(\theta,\beta)\in\Theta\times\mathcal{B}} \sum_{i=1}^{N} \log\left(\int_{\mathcal{A}} \prod_{t=1}^{T} f_{z|\alpha}(z_{it};\theta,\alpha_i,\psi) dF_\alpha(\alpha_i|z_{i0};\beta)\right). \tag{5.1}$$

In practice, the integral is not available on closed form and so is approximated using MC methods,

$$\left(\hat{\theta}_S,\hat{\beta}_S\right) = \arg\min_{(\theta,\beta)\in\Theta\times\mathcal{B}} \sum_{i=1}^{N} \log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{t=1}^{T} f_{z|\alpha}(z_{it};\theta,\alpha_s(\beta),\psi)\right), \tag{5.2}$$

where $\alpha_s(\beta)$, $s = 1,...,S$, are i.i.d. draws from $G(\cdot|z_{i0};\beta)$. This is the standard simulated MLE (SMLE).

As in the FE case, the SMLE is costly to compute if ψ is so: For a given value of (θ,β), the computation of the simulated likelihood requires NTS evaluations of ψ. Thus, unless NTS is "small", (θ,β) is low-dimensional, or ψ is fast to compute, the computation of the SMLE will be infeasible. Our interpolation scheme will also lead to substantial computational savings in an RE setting; the "indirect" ARE, where we first interpolate ψ and then plug it into the likelihood function, takes the form

$$\left(\hat{\theta}_{S,K},\hat{\beta}_{S,K}\right) = \arg\min_{(\theta,\beta)\in\Theta\times\mathcal{B}} \sum_{i=1}^{N} \log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{t=1}^{T} f_{z|\alpha}(z_{it};\theta,\alpha_s(\beta),\hat{\psi}_K)\right), \tag{5.3}$$

while the "direct" version is given by

$$\left(\hat{\theta}_{S,K},\hat{\beta}_{S,K}\right) = \arg\min_{(\theta,\beta)\in\Theta\times\mathcal{B}} \sum_{i=1}^{N} \log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{t=1}^{T} \hat{f}_{K,z|\alpha}(z_{it};\theta,\alpha_s(\beta),\psi)\right), \tag{5.4}$$

where $\hat{f}_{K,z|\alpha}$ denotes the interpolated version of $f_{K,z|\alpha}$. The discussion of the AFE carries over to the ARE with obvious modifications. In particular, derivatives of the simulated likelihood are easily computed due to interpolation thereby allowing for fast numerical computation of the optimization problems in (5.3) and (5.4).

Similar to the AFE, the ARE will suffer from additional biases due to interpolation and simulations. We propose to remove these by the split-panel Jackknife which is implemented the exact same way as for the ARE; in particular, we use the exact same draws $\alpha_1(\beta),....,\alpha_S(\beta)$ when computing $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$. The validity of this procedure is shown below.

5.1 Asymptotic Theory

5.2 Illustration: Buffer Stock Model

6 Higher Dimensional FE

1. Tensor product grids
2. Sparse grids

7 An Empirical Application to Danish Data

In this section, we fit the consumption model described in Section 2.1 to Danish administrative longitudinal register data using AFE. The empirical application is motivated by the increasing interest in allowing for ex ante heterogeneous agents in the standard work horse model of intertemporal consumption and wealth allocation. See, e.g., Alan and Browning (2010); Carroll, Slacalek, Tokuoka and White (2017); De Nardi and Fella (2017); Alan, Browning and Ejrnæs (2017); and Krueger, Mitman and Perri (2016). For simplicity and clarity of exposition, we focus on discount factor heterogeneity although such heterogeneity might capture heterogeneity across households in other dimensions. The degree of heterogeneity across households is of extreme importance for not only empirical work, but also for policy evaluations and recommendations. The empirical application of AFE is in turn an illustration of how flexible heterogeneity can be feasibly estimated in rich dynamic economic models using more than 200,000 households and almost 2,000,000 observations.

7.1 Data

We use high quality Danish administrative registers covering the entire population in the period 1987-1996.[8] All information is based on third party reports with little additional self-reporting. All self-reporting are moreover subject to possible auditing giving reliable longitudinal information on household characteristics, assets, liabilities and income.

Household income includes all monetary income net of all taxes, except any income related to ownership of financial assets. Transfers, such as child benefits and unemployment benefits, are also included to ensure that disposable income accurately measures the flow of resources available for consumption. Net wealth consists of stocks, bonds, bank deposits,

[8] We begin in 1987 to be able to consistently match individuals into couples, and we end with 1996 because the Danish wealth tax was abolished in this year. Information on, e.g., cars and boats where not collected in subsequent years leading to a break in the wealth measure from 1996 to 1997.

cars, boats, house value for home owners and mortgage deeds net of total liabilities. The house value is assessed by the tax authorities for tax purposes. Pension wealth is not observed in the registers and thus not included in the wealth measure.

Household consumption is not observed in the registers and is instead imputed using a simple budget approach, $C_t = \tilde{Y}_t - \Delta A_t$, where $\tilde{Y}_t = Y_t + r \cdot A_t$ is disposable income, A_t is end-of-period net wealth, r is the real rate of return, and ΔA_t thus proxies savings. A very similar imputation method is evaluated on Danish data in Browning and Leth-Petersen (2003) and found to produce a reasonable approximation. The resulting consumption measure will, however, e.g. include some durables such as home appliances. All variables are deflated with the official consumer price index.

We restrict attention to stable married or cohabiting couples in which the husband is between age 25 and 59. This is to mitigate issues regarding educational and retirement choices. To increase homogeneity of households, we restrict the spousal age difference to be no more than five years, and require that no one in the household ever becomes self-employed or are out of the labor market, and are neither a student nor retired. To limit the effect of errors in the imputation procedure on our estimates of time preference heterogeneity, we trim our sample from extreme observations and require that we have data for at least 5 years.[9] In total this leaves us with an unbalanced panel of 261,725 households observed in at most 9 time periods with a total of 1,966,741 household-time observations.

7.2 Calibrations

We fix some parameters of the model before turning to estimation of $\theta = (\gamma, \beta)$. These parameters are all reported in Table 7. Particularly, we choose an interest of $R = 1.03$ similar to the long run real return on 10 year Danish government bonds which over the period 1987-2007 was 3.8 percent. The same interest rate is used in e.g. Gourinchas and Parker (2002). Informally looking into the observed consumption behavior of households in debt we furthermore set the borrowing constraint to be binding at 30 percent of permanent income ($\lambda = 0.30$). Kaplan (2012) estimates an almost identical placement of the credit constraint using the PSID. Finally, we set the replacement rate in retirement to 90 percent ($\kappa = 0.9$) based on Danish Finance Ministry (2003) and assume that households retire at age 60 ($T = 59 - 25 + 1 = 35$) and dies at age 80 ($L = 55$). We fix the CRRA coefficient to $\rho = 1.5$.

Following the approach in Meghir and Pistaferri (2004), we estimate the transitory and permanent income shocks variances as, respectively, $\sigma^2_\xi = -\text{cov}(\Delta\epsilon_{it}, \Delta\epsilon_{i,t+1})$ and $\sigma^2_\psi =$

[9] Further details on the data are provided in Appendix C.

Table 7: Calibrated Parameters.

ρ	R	G	κ	σ_ψ	σ_ξ	π	T	L
1.5	1.03	Fig. 8	0.90	0.059	0.063	0.0	35	55

$\text{cov}(\Delta\epsilon_{it}, \sum_{k=0}^{2} \Delta\epsilon_{i,t-1+k})$, where ϵ_{it} is the residual for household i in period t from a regression of log household income on a full set of age and year dummies. The results are reported in Table 7. The income variances of Danish households are smaller than those typically estimated for the US. As argued in Jørgensen (2017), this is most likely due to *i)* a generous social welfare system, *ii)* progressive taxation, *iii)* a relatively high "minimum wage", and *iv)* register data is typically less noise compared to surveys typically used.

The growth in income is estimated for a given age as $G_t = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \Delta\epsilon_{it} + \frac{1}{2}\sigma_\psi^2\right)$ by re-arranging the income process. A smoothed growth rate $\tilde{G}_t$ is obtained using a third degree polynomial in age. The results are reported in Figure 8 in the Supplemental Material. Permanent income, P_{it}, is found by applying the Kalman filter on the time series of log income for each household.[10]

In line with the Monte Carlo study above, we use simple equally spaced tensor product grids for (m, β) when pre-computing the model for use in our AFE estimator. We restrict the domain of discount factors to the interval $[0.5, 1.05]$ in all estimations.

[10] We do not handle the inherent difficulties with the use of an estimated state variable in our estimation here. What ever bias may arise from this should also be present in the standard FE estimator as well as our AFE.

Table 8: Estimated Preferences.

	γ (homogeneous)		β (heterogeneous)			Time	Number
	Est.	SE	Med.†	Std.‡	Obj.	(mins.)	of sols.
Homogeneous	0.994	(0.001)	0.963	(0.0001)	0.955	0.12	1400
Fixed effects (FE)	0.481	(0.000)	0.935	[0.0517]	0.715	892.24	30811589
AFE, cubic spline interpolation							
$\hat{J}^{\S} = 55$ ($\eta = 10^{-4}$)	0.483	(0.000)	0.935	[0.0513]	0.716	0.57	990
$\hat{J}^{\S} = 120$ ($\eta = 10^{-5}$)	0.481	(0.000)	0.935	[0.0517]	0.715	1.18	2820
$J = 5$	0.438	(0.000)	0.910	[0.0759]	0.636	0.16	65
$J = 10$	0.473	(0.000)	0.936	[0.0463]	0.701	0.11	80
$J = 15$	0.462	(0.000)	0.934	[0.0542]	0.714	0.11	120
$J = 20$	0.480	(0.000)	0.934	[0.0483]	0.711	0.16	220
$J = 25$	0.485	(0.000)	0.935	[0.0508]	0.715	0.20	325
$J = 50$	0.482	(0.000)	0.935	[0.0511]	0.715	0.22	600
$J = 200$	0.481	(0.000)	0.935	[0.0518]	0.715	1.35	2200
AFE, linear interpolation							
$\hat{J}^{\S} = 70$ ($\eta = 10^{-4}$)	0.475	(0.000)	0.934	[0.0517]	0.714	0.74	1155
$\hat{J}^{\S} = 140$ ($\eta = 10^{-5}$)	0.480	(0.000)	0.935	[0.0519]	0.715	1.31	3710
$J = 5$	0.418	(0.000)	0.880	[0.1117]	0.615	0.26	65
$J = 10$	0.440	(0.000)	0.928	[0.0493]	0.682	0.19	90
$J = 15$	0.448	(0.000)	0.929	[0.0533]	0.698	0.21	150
$J = 20$	0.458	(0.000)	0.930	[0.0494]	0.701	0.20	200
$J = 25$	0.475	(0.000)	0.933	[0.0509]	0.707	0.33	450
$J = 50$	0.475	(0.000)	0.934	[0.0511]	0.713	0.37	850
$J = 200$	0.480	(0.000)	0.935	[0.0519]	0.715	0.29	2600
AFE, classification							
$\hat{J}^{\S} = 95$ ($\eta = 10^{-4}$)	0.481	(0.004)	0.933	[0.0530]	0.717	1.61	1520
$\hat{J}^{\S} = 190$ ($\eta = 10^{-5}$)	0.481	(0.004)	0.934	[0.0530]	0.716	4.62	4845
$J = 5$	0.374	(0.004)	0.913	[0.0553]	1.120	0.11	70
$J = 10$	0.429	(0.004)	0.928	[0.0545]	0.897	0.18	120
$J = 15$	0.530	(0.004)	0.932	[0.0537]	0.796	0.15	120
$J = 20$	0.499	(0.004)	0.934	[0.0534]	0.762	0.13	120
$J = 25$	0.506	(0.004)	0.935	[0.0531]	0.746	0.18	200
$J = 50$	0.487	(0.004)	0.938	[0.0530]	0.723	0.34	350
$J = 200$	0.481	(0.004)	0.934	[0.0530]	0.716	0.95	1200

Notes: Estimation results based on $N = 261,725$ households with $1,966,741$ household-year observations. For model and estimation details see discussion of Example 2 in Section 4. Remaining parameters are fixed at values in Table 7. Asymptotic standard errors clustered at the household level in brackets.

† Reported are the estimated homogeneous point estimate in the first row and all other rows report the estimated median (med.) of $\hat{\beta}_i$.

‡ Reported are the estimated asymptotic standard error (SE) on the homogeneous point estimate in the first row and all other rows report the estimated standard deviation (std.) of $\hat{\beta}_i$ in square brackets.

§ Denotes the estimated J using the method proposed in sub-section 2.4.3 with $l = 3$, $\Delta = 5$, $\gamma = 0.5$ and a tolerance of η. Timings reported in these columns are the total estimation time of finding $\hat{J}$ and subsequent estimation of γ using $\hat{J}$ nodes.

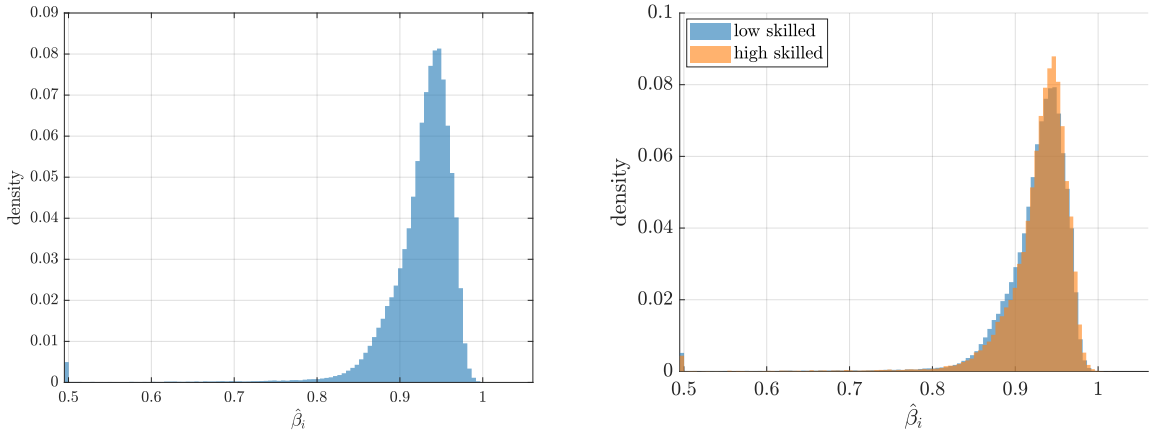
7.3 Estimation Results

The estimation results are presented in Table 8. When both parameters are restricted to be homogeneous, we estimate γ to be around 1 and β to be around 0.96. The discount factor is well within the range typically found and the retirement parameter, γ, yields a marginal propensity to consume out of wealth in retirement close to what is estimated in Gourinchas and Parker (2002). They estimate the marginal propensity to consume in retirement to be around 7% in the PSID while our homogeneous estimation results suggest a marginal propensity to consume in retirement of around 6.8%.[11]

The FE estimate of γ, where we allow β_i to be household-specific, reduces to around 0.48 while the median discount factor is estimated to be around 0.935 and the standard deviation of the distribution is 0.052. The estimated distribution of discount factors are shown in Figure 6. We note that while the distribution is left skewed, almost all the mass of the distribution is within 0.80 and 1.00. Furthermore, the estimated distributions are very similar across educational groups as seen in the right panel of Figure 6.[12] Households with more education tend to be relatively less impatient with the distribution of discount factors shifted slightly to the right.

Figure 6: Distribution of Estimated Discount Factors, $\hat{\beta}_i$.

(a) All households. (b) Split by Education.



Notes: The right panel reports the estimated distribution of heterogeneous discount factors split by educational attainment. Households are classified as high skilled if either member holds at least a bachelor degree (70.784 households are classified as high skilled).

In Table 8, below the FE estimates, we report the estimation results for various implementations of our proposed AFE estimator. Particularly, we show results when using *i)*

[11] We calculate the marginal propensity to consume in retirement as $\hat{\gamma} \cdot \frac{1-R^{-1}(\hat{\beta}R)^{1/\rho}}{1-[R^{-1}(\hat{\beta}R)^{1/\rho}]^{L-T}} = 0.068$ based on the formula for consumption in retirement in eq. (**??**).

[12] Households are classified as high skilled if either member holds at least a bachelor degree.

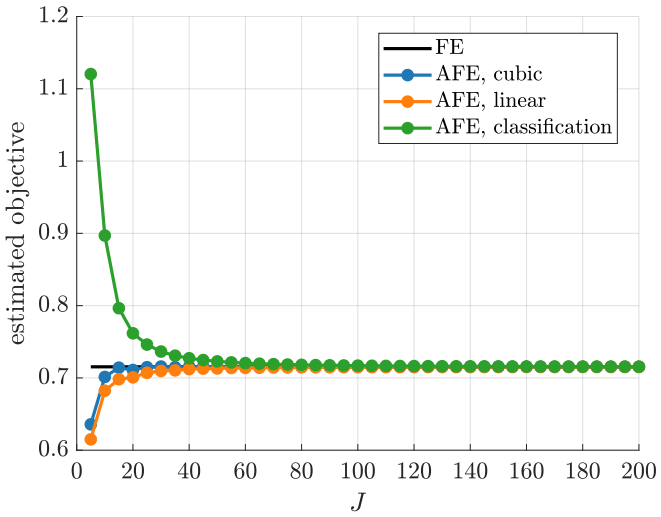
cubic spline interpolation, *ii)* linear interpolation, and *iii)* classification. We show results both for an a priori chosen number of pre-computation nodes in each dimension, J, and when choosing the number of pre-computation nodes using the approach proposed in sub-section 2.4.3. All computations were done on a high-powered computer system using 56 cores of 2.00 GHz.

Across all implementations using as little as 25 nodes seems to give reasonably similar estimates as the FE. In these cases our AFE estimator is almost or more than 3,000 times faster than the FE estimator. The main explanation is that while the FE estimator requires more than 30 million solutions of the dynamic programming problem, our AFE estimators require a few hundred. Figure 7 illustrates how the AFE objective functions converge towards the FE objective function as J increases. The rate of convergence clearly increases in the smoothness of the interpolant.

Choosing the number of pre-computation nodes, J, as proposed in sub-section 2.4.3 with a tolerance of $\eta = 10^{-4}$, implies 55 nodes when using spline interpolation, 70 when using linear interpolation, and 95 when using classification. Across all implementations, the AFE results are now very close to the FE results when choosing J by this data driven approach. The speed-ups also remain size-able. With spline interpolation, the fastest of the three interpolation schemes in our setting, the AFE estimator is more than $1,500$ times faster than the standard FE estimator. To underline the scope of this difference in speed, note that if it takes 10 *minutes* to use our AFE estimator, it will take 10 *days* to use the standard FE estimator.[13]

[13]Lowering the tolerance to $\eta = 10^{-5}$ more or less double the required number of nodes cutting the speed-up factor in half. For a given tolerance, the speed-ups we report here can, however, be seen as lower bounds as we could further optimize the AFE estimator by both using non-equally spaced grids and relatively more nodes in the m-dimension, which would not require additional solutions of the dynamic programming problem, but would increase the precision of the interpolant.

Figure 7: Convergence of AFE to FE.



Notes: The figure illustrates the convergence of the AFE objective function to the FE objective function. Each dot represents the objective function when re-estimating all parameters using a given J.

8 Conclusion

To be added.

References

Alan, S. and M. Browning (2010): "Estimating Intertemporal Allocation Parameters using Synthetic Residual Estimation," *The Review of Economic Studies*, 77(4), 1231–1261.

Alan, S., M. Browning and M. Ejrnæs (2017): "Income and Consumption: a Micro Semi-structural Analysis with Pervasive Heterogeneity," *forthcoming, Journal of Political Economy*.

Andersen, S., G. W. Harrison, M. I. Lau and E. E. Rutström (2008): "Eliciting Risk and Time Preferences," *Econometrica*, 76(3), 583–618.

Andreoni, J. and C. Sprenger (2012): "Risk Preferences Are Not Time Preferences," *The American Economic Review*, 102(7), 3357–3376.

Bajari, P., J. T. Fox and S. P. Ryan (2007): "Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients," *The American Economic Review*, 97(2), 459–463.

Barsky, R. B., F. T. Juster, M. S. Kimball and M. D. Shapiro (1997): "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study," *The Quarterly Journal of Economics*, 112(2), 537–579.

Beetsma, R. M. W. J. and P. C. Schotman (2001): "Measuring Risk Attitudes in a Natural Experiment: Data from the Television Game Show Lingo," *The Economic Journal*, 111(474), 821–848.

Bester, C. A. and C. B. Hansen (2015): "Grouped effects estimators in fixed effects models," *Journal of Econometrics*.

Birkhoff, G. and A. Priver (1967): "Hermite Interpolation Errors for Derivatives," *Journal of Mathematics and Physics*, 46(1-4), 440–447.

Bonhomme, S., T. Lamadon and E. Manresa (2017): "Discretizing Unobserved Heterogeneity," Discussion paper.

Bonhomme, S. and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83(3), 1147–1184.

Browning, M. and S. Leth-Petersen (2003): "Imputing Consumption from Income and Wealth Information," *The Economic Journal*, 113(488), F282–F301.

Cagetti, M. (2003): "Wealth Accumulation Over the Life Cycle and Precautionary Savings," *Journal of Business & Economic Statistics*, 21(3), 339–353.

Carroll, C., J. Slacalek, K. Tokuoka and M. N. White (2017): "The distribution of wealth and the marginal propensity to consume," *Quantitative Economics*, 8(3), 977–1020.

Carroll, C. D. (1992): "The buffer-stock theory of saving: Some macroeconomic evidence," *Brookings Papers on Economic Activity*, 2, 61–156.

——— (1997): "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis," *The Quarterly Journal of Economics*, 112(1), 1–55.

——— (2006): "The method of endogenous gridpoints for solving dynamic stochastic optimization problems," *Economics Letters*, 91(3), 312–320.

——— (2012): "Theoretical Foundations of Buffer Stock Saving," Working Paper, http://www.econ2.jhu.edu/people/ccarroll/papers/BufferStockTheory/.

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

Ciliberto, F. and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

Coller, M. and M. B. Williams (1999): "Eliciting Individual Discount Rates," *Experimental Economics*, 2(2), 107–127.

Danish Finance Ministry (2003): "Ældres sociale vilkår (in Danish)," Discussion paper.

De Nardi, M. and G. Fella (2017): "Saving and wealth inequality," *Review of Economic Dynamics*, 26, 280–300.

Deaton, A. (1991): "Saving and liquidity constraints," *Econometrica*, 59(5), 1221–1248.

Deaton, A. (1992): *Understanding Consumption*. Oxford University Press.

Dhaene, G. and K. Jochmans (2015): "Split-panel Jackknife Estimation of Fixed-effect Models," *The Review of Economic Studies*, 82(3), 991–1030.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp and G. G. Wagner (2011): "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences," *Journal of the European Economic Association*, 9(3), 522–550.

Fernández-Villaverde, J., J. F. Rubio-Ramírez and M. S. Santos (2006): "Convergence properties of the likelihood of computed dynamic models," *Econometrica*, 74(1), 93–119.

Finke, M. S. and S. J. Huston (2013): "Time preference and the importance of saving for retirement," *Journal of Economic Behavior & Organization*, 89, 23–34.

Fox, J. T., K. i. Kim, S. P. Ryan and P. Bajari (2011): "A simple estimator for the distribution of random coefficients," *Quantitative Economics*, 2(3), 381–418.

Fox, J. T., K. i. Kim and C. Yang (2016): "A simple nonparametric approach to estimating the distribution of random coefficients in structural models," *Journal of Econometrics*, 195(2), 236–254.

French, E. and J. B. Jones (2011): "The Effects of Health Insurance and Self-Insurance on Retirement Behavior," *Econometrica*, 79(3), 693–732.

Gourinchas, P.-O. and J. A. Parker (2002): "Consumption over the life cycle," *Econometrica*, 70(1), 47–89.

Guiso, L. and M. Paiella (2008): "Risk Aversion, Wealth, and Background Risk," *Journal of the European Economic Association*, 6(6), 1109–1150.

Hahn, J. and G. Kuersteiner (2011): "Bias reduction for dynamic nonlinear panel models with fixed effects," *Econometric Theory*, 27(6), 1152–1191.

Hahn, J. and H. R. Moon (2010): "Panel data model with finite number of multiple equilibria," *Econometric Theory*, 26(3), 863–881.

Hahn, J. and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Han, X. (????): "A Two-Step Estimator for Structural Models Using Approximation," .

Heckman, J. (1981): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process," in *Structural Analysis of Discrete Panel Data with Econometric Applications,*, ed. by C. F. Manski and D. McFadden, pp. 179–195. MIT Press, Cambridge, MA.

Heckman, J. and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, pp. 271–320.

Holt, C. A. and S. K. Laury (2005): "Risk Aversion and Incentive Effects: New Data without Order Effects," *The American Economic Review*, 95(3), 902–904.

Howell, G. W. (1991): "Derivative error bounds for Lagrange interpolation: An extension of Cauchys bound for the error of Lagrange interpolation," *Journal of Approximation Theory*, 67(2), 164–173.

Jørgensen, T. H. (2017): "Life-Cycle Consumption and Children: Evidence from a Structural Estimation," *Oxford Bulletin of Economics and Statistics*, 79(5), 717–746.

Judd, K. (1998): *Numerical Methods in Economics*. MIT Press.

Kamakura, W. A. (1991): "Estimating flexible distributions of ideal-points with external analysis of preferences," *Psychometrika*, 56(3), 419–431.

Kaplan, G. (2012): "Inequality and the life cycle," *Quantitative Economics*, 3(3), 471–525.

Keane, M. P. and K. I. Wolpin (1994): "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *The Review of Economics and Statistics*, 76(4), 648.

Kristensen, D. and B. Salanié (2017): "Higher-order properties of approximate estimators," *Journal of Econometrics*, 198(2), 189–208.

Kristensen, D. and Y. Shin (2012): "Estimation of dynamic models with nonparametric simulated maximum likelihood," *Journal of Econometrics*, 167(1), 76–94.

Krueger, D., K. Mitman and F. Perri (2016): "Chapter 11 - Macroeconomics and Household Heterogeneity," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, vol. 2, pp. 843–921. Elsevier, DOI: 10.1016/bs.hesmac.2016.04.003.

Low, H., C. Meghir and L. Pistaferri (2010): "Wage Risk and Employment Risk over the Life Cycle," *American Economic Review*, 100(4), 1432–1467.

Meghir, C. and L. Pistaferri (2004): "Income variance dynamics and heterogeneity," *Econometrica*, 72(1), 1–32.

Nevo, A., J. L. Turner and J. W. Williams (2016): "Usage-Based Pricing and Demand for Residential Broadband," *Econometrica*, 84(2), 411–443.

Zhou, S. and D. A. Wolfe (2000): "On Derivative Estimation in Spline Regressions," *Statistica Sinica*, (10), 93–108.

A Lemmas and Proofs

A.1 Lemmas

The following limit results will be used in our main proofs:

Lemma A.1. *Let* $a_{it}(\phi,\psi) = a(z_{it};\phi,\psi)$ *for* $\phi \in \Phi$ *where* $\Phi \subseteq \mathbb{R}^{d_\phi}$ *is compact and convex, where* $\{z_{i,t}\}$ *(i) Suppose that* $a_{it}(\phi,\psi)$ *is* L_p-Lipshitz continuous w.r.t. (ϕ,ψ) with $p > 4$ and $\left\|\hat{\psi} - \psi_0\right\| = o_P(1)$. Then

$$\max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \left\| \frac{1}{T}\sum_{t=1}^{T} \left\{ a_{it}\left(\phi,\hat{\psi}\right) - E\left[a_{it}(\phi,\psi_0)\right] \right\} \right\| = o_P(1).$$

(ii) Furthermore, if $p > 10(1+p_0)/(1-10v)$, where $p_0 \geq (d_\phi+4)/2$, and $\left\|\hat{\psi} - \psi_0\right\| = O_P(T^{-\rho})$, with $\rho > 0$ defined in Assumption 1, then

$$\max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \left\| \frac{1}{T}\sum_{t=1}^{T} \left\{ a_{it}\left(\phi,\hat{\psi}\right) - E\left[a_{it}(\phi,\psi_0)\right] \right\} \right\| = O_P\left(T^{-\rho}\right).$$

Proof. Write

$$\max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \left\| \frac{1}{T}\sum_{t=1}^{T} a_{it}\left(\phi,\hat{\psi}\right) - E\left[a_{it}(\phi,\psi_0)\right] \right\| \leq \max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \frac{1}{T}\sum_{t=1}^{T} \left\| a_{it}\left(\phi,\hat{\psi}\right) - a_{it}(\phi,\psi_0) \right\| + \max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \left\| \frac{1}{T}\sum_{t=1}^{T} a_{it}(\phi,\psi_0) - E\left[a_{it}(\phi,\psi_0)\right] \right\|, \tag{A.1}$$

where the first term satisfies, with $B_\xi(z)$ denoting the Lipschitz "coefficient" of $\xi(z;\phi,\psi)$,

$$\max_{i=1,\ldots,N} \sup_{\phi\in\Phi} \frac{1}{T}\sum_{t=1}^{T} \left\| a_{it}\left(\phi,\hat{\psi}\right) - a_{it}(\phi,\psi_0) \right\| \leq \left\{ \max_{i=1,\ldots,N} \left\| \frac{1}{T}\sum_{t=1}^{T} B_\xi(z_{it}) - E\left[B_\xi(z_{it})\right] \right\| \right\} \times \left\|\hat{\psi} - \psi_0\right\| + \max_{i=1,\ldots,N} E\left[B_\xi(z_{it})\right] \left\|\hat{\psi} - \psi_0\right\|. \tag{A.2}$$

The first part of the lemma now follows by applying Lemma 1 of HK to the second term of eq. (A.1) and the first term of eq. (A.2) together with the first convergence condition imposed on $\hat{\psi}$ in the lemma and the fact that $\max_{i=1,\ldots,N} E\left[B_\xi(z_{it})\right] = O(1)$. The second part is obtained by applying Lemma 2 instead of Lemma 1 of HK together with the strengthened convergence condition imposed on $\hat{\psi}$. □

Lemma A.2. *Let* $a_{it}(\phi,\psi) = a(z_{it};\phi,\psi)$ *and* $b_{it}(\phi,\psi) = b(z_{it};\phi,\psi)$ *for* $\phi \in \Phi$ *where*

$\Phi \subseteq \mathbb{R}^{d_\phi}$ *is compact and convex.* Then

$$\text{(i)}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} a_{it}(\phi,\psi_0) \to^p E_\infty\left[a_{it}(\phi,\psi_0)\right],$$

$$\text{(ii)}\ \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{a_{it}(\phi,\psi_0) - E\left[a_{it}(\phi,\psi_0)\right]\right\} \to^d N\left(0, Var_\infty\left(a_{it}(\phi,\psi_0)\right)\right)$$

$$\text{(iii)}\frac{1}{NT}\sum_{i=1}^{N}\sum_{s,t=1}^{T}\left\{a_{is}(\phi,\psi_0) - E\left[a_{is}(\phi,\psi_0)\right]\right\}\left\{b_{it}(\phi,\psi_0) - E\left[b_{it}(\phi,\psi_0)\right]\right\} \to^p Cov_\infty\left(a_{it}(\phi,\psi_0), b_{it}(\phi,\psi_0)\right.$$

where $E_\infty\left[a_{it}(\phi,\psi_0)\right] := \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} E\left[a_{it}(\phi,\psi_0)\right]$ and similar for the other expectations.

A.2 Proof of Theorem 3.2

We first show that the estimators are consistent:

Theorem A.1. *Under Assumptions 1-2(i),* $\left\|\hat{\theta} - \theta_0\right\| = o_P(1)$. *Moreover, w.p.a.1., there exists functions* $\hat{\alpha}_i\left(\theta,\hat{\psi}\right)$ *and* $\alpha_{0,i}(\theta,\psi_0)$ *solving*

$$\bar{s}_i^{(\alpha)}\left(\theta,\hat{\alpha}_i\left(\theta,\hat{\psi}\right),\hat{\psi}\right) = 0,\ S_i^{(\alpha)}\left(\theta,\alpha_{0,i}(\theta,\psi_0),\psi_0\right) = 0, \tag{A.3}$$

respectively for θ *in a neighbourhood of* θ_0. *The functions are continuously differentiable and satisfy*

$$\sup_{\|\theta-\theta_0\|<\epsilon}\left\|\hat{\alpha}\left(\theta,\hat{\psi}\right) - \alpha_0(\theta,\psi_0)\right\|_\infty = o_P(1),\ \sup_{\|\theta-\theta_0\|<\epsilon}\left\|\frac{\partial\hat{\alpha}\left(\theta,\hat{\psi}\right)}{\partial\theta'} - \frac{\partial\alpha_0(\theta,\psi_0)}{\partial\theta'}\right\|_\infty = o_P(1).$$

If furthermore Assumption 2(ii) hold, then $\sup_{\|\theta-\theta_0\|<\epsilon}\left\|\hat{\alpha}\left(\theta,\hat{\psi}\right) - \alpha_0(\theta,\psi_0)\right\|_\infty = o_P\left(T^{-\rho}\right)$ *with* ρ *defined in Assumption 1-2(ii).*

Proof of Theorem A.1. By Lemma A.1, $\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}}\left\|\bar{q}\left(\theta,\alpha,\hat{\psi}\right) - Q(\theta,\alpha,\psi_0)\right\|_\infty = o_P(1)$. The first part of the theorem now follows by the exact same arguments as the ones in Theorems 3 and 4 in HK; these arguments also yield

$$\sup_{\theta\in\mathcal{B}(\theta_0,\epsilon)}\left\|\hat{\alpha}\left(\theta,\hat{\psi}\right) - \alpha_0(\theta,\psi_0)\right\|_\infty = o_P(1), \tag{A.4}$$

for some $\epsilon > 0$, where $\mathcal{B}(\theta_0,\epsilon) = \{\theta\in\Theta : \|\theta-\theta_0\| < \epsilon\}$. To show the second part, first observe that since $\alpha_{0,i}(\theta_0,\psi_0)$ is situated in the interior of $\mathcal{A}$, it must satisfy the second equation of (A.3). Moreover, by Condition 6 in HK, $H_i^{(\alpha,\alpha)}(\theta,\alpha,\psi_0)$ exists, has full rank at $(\theta_0,\alpha_{0,i}(\theta_0,\psi_0))$ and is continuous w.r.t. (θ,α). It then follows by the Implicit Function Theorem that there exists a function $\alpha_{0,i}(\theta,\psi_0)$ satisfying $S_i^{(\alpha)}\left(\theta,\alpha_{0,i}(\theta,\psi_0),\psi_0\right) = 0$

for $\theta \in \mathcal{B}(\theta_0, \epsilon)$ (potentially after choosing a smaller value for $\epsilon > 0$). Given that the eigenvalues of $H_i^{(\alpha,\alpha)}(\theta_0, \psi_0)$ are assumed to be bounded away from zero uniformly over $i = 1, ..., N$, $\alpha_{0,i}(\theta, \psi_0)$ will be continuously differentiable w.r.t. θ uniformly over $i = 1, ..., N$. Next, due to (A.4), $\hat{\alpha}_i\left(\theta, \hat{\psi}\right)$, $\theta \in \mathcal{B}(\theta_0, \epsilon)$, will also be situated in the interior of $\mathcal{A}$ w.p.a.1 uniformly over i. Thus, it must satisfy $\bar{s}_i^{(\alpha)}\left(\hat{\theta}, \hat{\alpha}_i\left(\hat{\theta}, \hat{\psi}\right), \hat{\psi}\right) = 0$. By another application of Lemma A.1(i), $\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}} \left\|\bar{h}^{(\alpha,\alpha)}\left(\theta, \alpha, \hat{\psi}\right) - H^{(\alpha,\alpha)}(\theta, \alpha, \psi_0)\right\|_\infty = o_P(1)$ where $H_i^{(\alpha)}(\theta, \alpha, \psi_0)$ is continuous w.r.t. (θ, α) uniformly over i. This combined with the first part of the theorem yields $\left\|\bar{h}^{(\alpha,\alpha)}\left(\hat{\theta}, \hat{\alpha}\left(\hat{\theta}, \hat{\psi}\right), \hat{\psi}\right) - H^{(\alpha,\alpha)}(\theta_0, \alpha_0(\theta_0, \psi_0), \psi_0)\right\|_\infty = o_P(1)$. It therefore also holds w.p.a.1 that there exists a function $\hat{\alpha}_i\left(\theta, \hat{\psi}\right)$ so that eq. (A.3) holds for $\theta \in \mathcal{B}(\theta_0, \epsilon)$. For any $\theta \in \mathcal{B}(\theta_0, \epsilon)$, the following arguments are valid: First, by the mean value theorem,

$$0 = \bar{s}_i^{(\alpha)}\left(\theta, \alpha_{0,i}(\theta, \psi_0), \hat{\psi}\right) + \bar{h}_i^{(\alpha,\alpha)}\left(\theta, \bar{\alpha}_i(\theta), \hat{\psi}\right)\left\{\hat{\alpha}_i\left(\theta, \hat{\psi}\right) - \alpha_{0,i}(\theta, \psi_0)\right\} \tag{A.5}$$

where $\bar{\alpha}_i(\theta)$ is situated on the line segment connecting $\alpha_{0,i}\left(\theta, \hat{\psi}\right)$ and $\alpha_{0,i}(\theta, \psi_0)$. By Lemma A.1(i),

$$\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}} \left\|\bar{s}_i^{(\alpha)}\left(\theta, \alpha_0(\theta, \psi_0), \hat{\psi}\right)\right\| = o_P(1). \tag{A.6}$$

Moreover, from the earlier part of the proof, we know that all eigenvalues of $\bar{h}_i^{(\alpha,\alpha)}\left(\theta, \bar{\alpha}_i(\theta), \hat{\psi}\right)$ are bounded away from zero uniformly over i w.p.a.1. Thus, there exists $c > 0$ so that w.p.a.1,

$$\left\|\bar{h}_i^{(\alpha,\alpha)}\left(\theta, \hat{\alpha}_i\left(\theta, \hat{\psi}\right), \hat{\psi}\right)\left\{\hat{\alpha}_i\left(\theta, \hat{\psi}\right) - \alpha_{0,i}(\theta, \psi_0)\right\}\right\| \geq c\left\|\hat{\alpha}_i\left(\theta, \hat{\psi}\right) - \alpha_{0,i}(\theta, \psi_0)\right\|. \tag{A.7}$$

Eqs. (A.5)-(A.7) combined show uniform consistency of $\hat{\alpha}_i(\theta)$. Next, by taking derivatives w.r.t. θ on both sides of $S_i^{(\alpha)}(\theta, \alpha_i(\theta, \psi_0), \psi_0) = 0$, the expression for $\partial\alpha_{0,i}(\theta, \psi_0)/(\partial\theta)$ is obtained and similar for $\partial\hat{\alpha}_i\left(\theta, \hat{\psi}\right)/(\partial\theta)$. The proof of uniform consistency of $\partial\hat{\alpha}\left(\theta, \hat{\psi}\right)/(\partial\theta)$ now follows along the same lines as the above analysis of $\hat{\alpha}\left(\theta, \hat{\psi}\right)$ and so is left out. Finally, if we impose Assumption 2(ii), we obtain from Lemma A.1(ii) that eq. (A.6) can be strengthened to $\sup_{(\theta,\alpha)\in\Theta\times\mathcal{A}} \left\|\bar{s}_i^{(\alpha)}\left(\theta, \alpha_0\left(\theta, \hat{\psi}\right), \hat{\psi}\right)\right\| = o_P(T^{-\rho})$ which together with eqs. (A.5) and (A.7) yield the final part of the Theorem. □

Next, we derive the expansion in Theorem 3.1 where the leading terms are sample averages of

$$u_{i,t}(\psi) = -H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1} s_{0,i,t}^{(\alpha)}(\psi) \in \mathbb{R}^{d_\alpha}, \tag{A.8}$$

$$v_{i,t}(\psi) = -H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\left\{h_{0,i,t}^{(\alpha,\alpha)}(\psi) - H_{0,i}^{(\alpha,\alpha)}(\psi_0) + \frac{1}{2}\sum_{k=1}^{d_\alpha} u_{i,t,k}(\psi) G_{0,i,k}^{(\alpha,\alpha,\alpha_k)}(\psi_0)\right\} \in \mathbb{R}^{d_\alpha \times d_\alpha}, \tag{A.9}$$

with $u_{i,t,k}(\psi)$ being the kth element of $u_{i,t}(\psi)$, $H_{0,i}^{(\alpha,\alpha)}(\psi) = H_{0,i}^{(\alpha,\alpha)}(\theta_0, \alpha_{0,i}(\theta_0,\psi),\psi)$ and similar for other fucntions. Here,

$$\begin{aligned} s_{i,t}^{(\alpha)}(\theta,\alpha,\psi) &= \frac{\partial q_{i,t}(\theta,\alpha,\psi)}{\partial\alpha},\ h_{i,t}^{(\alpha,\alpha)}(\theta,\alpha,\psi) = \frac{\partial^2 q_{i,t}(\theta,\alpha,\psi)}{\partial\alpha\partial\alpha'}, \\ g_{i,t}^{(\alpha,\alpha,\alpha_k)}(\theta,\alpha,\psi) &= \frac{\partial^3 q_{i,t}(\theta,\alpha,\psi)}{\partial\alpha\partial\alpha'\partial\alpha_k}. \end{aligned}$$

Note here that if ψ depends on α, then the above functions involve derivatives of ψ w.r.t. α. Note that all population moments in the above definitions are evaluated at ψ_0 and only the random terms depend on $\hat{\psi}$. Setting $\hat{\psi} = \psi_0$ in the above, it is easily seen that $u_{i,t}(\psi_0)$ and $v_{i,t}(\psi_0)$ both have zero mean and so Lemma A.1(ii) can be used to obtain their uniform rates.

Proof of Theorem 3.1. In the following write $\alpha_{0,i} = \alpha_i(\theta_0,\psi_0)$, $\hat{\alpha}_i = \hat{\alpha}_i\left(\theta_0,\hat{\psi}\right)$.We proceed as in the proof of Lemma A4 of Newey and Smith (2004): First, by a second order Taylor expansion of the first-order condition (A.3) around $\alpha_{0,i}$,

$$0 = \bar{s}_{0,i}^{(\alpha)}\left(\hat{\psi}\right) + \bar{h}_{0,i}^{(\alpha,\alpha)}\left(\hat{\psi}\right)(\hat{\alpha}_i - \alpha_{0,i}) + \frac{1}{2}\sum_{k=1}^{d_\alpha}(\hat{\alpha}_{i,k} - \alpha_{0,i,k})'\,\bar{g}_i^{(\alpha,\alpha,\alpha_k)}\left(\bar{\alpha}_i,\hat{\psi}\right)(\hat{\alpha}_i - \alpha_{0,i})$$

where $\bar{\alpha}_i$ lies on the line segment connecting $\alpha_{0,i}$ to $\hat{\alpha}_i$. Add and subtract $H_{0,i}^{(\alpha,\alpha)}(\psi_0)(\hat{\alpha}_i - \alpha_{0,i})$, multiply through with $H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}$ and then rearrange to obtain

$$\begin{aligned} \hat{\alpha}_i - \alpha_{0,i} = \bar{u}_i\left(\hat{\psi}\right) &- H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\left\{\bar{h}_{0,i}^{(\alpha,\alpha)}\left(\hat{\psi}\right) - H_{0,i}^{(\alpha,\alpha)}(\psi_0)\right\}(\hat{\alpha}_i - \alpha_{0,i}) \\ &- \frac{1}{2}\sum_{k=1}^{d_\alpha}(\hat{\alpha}_{i,k} - \alpha_{0,i,k})'\,H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\,\bar{g}_i^{(\alpha,\alpha,\alpha_k)}\left(\bar{\alpha}_i,\hat{\psi}\right)(\hat{\alpha}_i - \alpha_{0,i}). \end{aligned}$$

Combining the convergence rate result in Theorem A.1 with Lemma A.1,

$$\left\|\bar{g}_i^{(\alpha,\alpha,\alpha_k)}\left(\bar{\alpha}_i,\hat{\psi}\right) - G_{i,0}^{(\alpha,\alpha,\alpha_k)}(\psi_0)\right\| \leq \left\{\frac{1}{T}\sum_{t=1}^{T}B_g(z_{i,t})\right\}\left\{\|\bar{\alpha} - \alpha_0\|_\infty + \left\|\hat{\psi} - \psi_0\right\|\right\} = \bar{O}_P\left(T^{-\rho}\right),$$

and so

$$\frac{1}{2}\sum_{k=1}^{d_\alpha}(\hat{\alpha}_{i,k} - \alpha_{0,i,k})'\,H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\left\{\bar{g}_i^{(\alpha,\alpha,\alpha_k)}\left(\bar{\alpha}_i,\hat{\psi}\right) - G_{i,0}^{(\alpha,\alpha,\alpha_k)}(\psi_0)\right\}(\hat{\alpha}_i - \alpha_{0,i}) = \bar{O}_P\left(T^{-3\rho}\right),$$

which in turn implies

$$\begin{aligned} \hat{\alpha}_i - \alpha_{0,i} = \bar{u}_i\left(\hat{\psi}\right) &- H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\left\{\bar{h}_{0,i}^{(\alpha,\alpha)}\left(\hat{\psi}\right) - H_{0,i}^{(\alpha,\alpha)}(\psi_0)\right\}(\hat{\alpha}_i - \alpha_{0,i}) \\ &- \frac{1}{2}\sum_{k=1}^{q}(\hat{\alpha}_{i,k} - \alpha_{0,i,k})'\,H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\,G_{0,i}^{(\alpha,\alpha,\alpha_k)}(\psi_0)(\hat{\alpha}_i - \alpha_{0,i}) + \bar{O}_P\left(T^{-3\rho}\right). \end{aligned}$$

Repeated use of Theorem A.1 and Lemma A.1 reveals that the two last terms on the right-hand side are $\bar{O}_P\left(T^{-2\rho}\right)$, and so $\hat{\alpha}_i - \alpha_{0,i} = \bar{u}_i\left(\hat{\psi}\right) + \bar{O}_P\left(T^{-2\rho}\right)$. Substituting the right-hand side of this last expression into the above display yields eq. (3.5). □

We now use Theorem 3.1 to develop an expansion of $\hat{\theta}$ w.r.t. $\hat{\alpha}_i$, $i = 1, ..., N$. This will involve the following terms:

$$s_{i,t}^{(\theta)}(\theta,\alpha,\psi) = \frac{\partial q_{i,t}(\theta,\alpha,\psi)}{\partial\theta},\ h_{i,t}^{(\theta,\theta)}(\theta,\alpha,\psi) = \frac{\partial^2 q_{i,t}(\theta,\alpha,\psi)}{\partial\theta\partial\theta'},$$

$$\begin{aligned} S_i^{(\theta)}(\theta,\alpha,\psi) &= \frac{\partial Q_i(\theta,\alpha,\psi)}{\partial\theta},\ H_i^{(\theta,\theta)}(\theta,\alpha,\psi) = \frac{\partial^2 Q_i(\theta,\alpha,\psi)}{\partial\theta\partial\theta'}, \\ g_{i,t}^{(\theta,\alpha,\alpha_k)}(\theta,\alpha,\psi) &= \frac{\partial^3 q_{i,t}(\theta,\alpha,\psi)}{\partial\theta\partial\theta'\partial\alpha_k},\ G_i^{(\theta,\alpha,\alpha_k)}(\theta,\alpha,\psi) = \frac{\partial^3 Q_i(\theta,\alpha,\psi)}{\partial\theta\partial\theta'\partial\alpha_k}, \end{aligned}$$

and similar for other partial derivatives. Note here that if ψ depends on θ, then the above functions involve derivatives of ψ w.r.t. θ. With this notation, we obtain the following expansion of $\hat{\theta}$ w.r.t. the FE estimators:

Theorem A.2. *Under Assumptions 1-2,*

$$\hat{\theta} - \theta_0 = \Phi_N\left(\hat{\psi}\right) + B_N\left(\hat{\psi}\right)/T + O_P\left(T^{-3\rho}\right), \tag{A.10}$$

where

$$\Phi_N\left(\hat{\psi}\right) = H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{1}{N}\sum_{i=1}^{N}\left\{\bar{s}_{0,i}^{(\theta)}\left(\hat{\psi}\right) + H_{0,i}^{(\theta,\alpha)}(\psi_0)\,\bar{u}_i\left(\hat{\psi}\right)\right\} \tag{A.11}$$

and

$$\begin{aligned} B_N\left(\hat{\psi}\right) &= H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{T}{N}\sum_{i=1}^{N}\left\{\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right) - H_{0,i}^{(\theta,\alpha)}(\psi_0)\right\}\bar{u}_i\left(\hat{\psi}\right) && \text{(A.12)} \\ &\quad + H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{T}{N}\sum_{i=1}^{N}H_{0,i}^{(\theta,\alpha)}(\psi_0)\,\bar{v}_i\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right) && \text{(A.13)} \\ &\quad + \frac{1}{2}H_0^{(\theta,\theta)}(\psi_0)^{-1}\sum_{k=1}^{d_\alpha}\frac{T}{N}\sum_{i=1}^{N}G_{0,i}^{(\theta,\alpha,\alpha_k)}(\psi_0)\,\bar{u}_i\left(\hat{\psi}\right)\bar{u}_{i,k}\left(\hat{\psi}\right). \end{aligned}$$

This expansion generalizes the one found in eq. (7) of HK to allow for the presence of a first-step estimator/approximator $\hat{\psi}$. The discussion following Theorem 3.1 also applies here.

Proof of Theorem A.2. By assumption, θ_0 lies in the interior of Θ. It then follows from Theorem A.1, that $\hat{\theta}$ is also situated in the interior w.p.a.1 and so the following first-order

condition is valid,

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial \bar{q}_i\left(\theta,\hat{\alpha}_i\left(\theta\right),\hat{\psi}\right)}{\partial\theta}\right|_{\theta=\hat{\theta}} \\ &= \frac{1}{N}\sum_{i=1}^{N}\bar{s}_i^{(\theta)}\left(\hat{\theta},\hat{\alpha}_i\left(\hat{\theta}\right),\hat{\psi}\right)+\frac{1}{N}\sum_{i=1}^{N}\bar{s}_i^{(\alpha)}\left(\hat{\theta},\hat{\alpha}_i\left(\hat{\theta}\right),\hat{\psi}\right)\frac{\partial\hat{\alpha}_i\left(\hat{\theta},\hat{\psi}\right)}{\partial\theta} \\ &= \frac{1}{N}\sum_{i=1}^{N}\bar{s}_i^{(\theta)}\left(\hat{\theta},\hat{\alpha}_i\left(\hat{\theta}\right),\hat{\psi}\right), \end{aligned}$$

where the second equality uses the chain rule and the third one Theorem A.1. Next, by the mean-value theorem,

$$0 = \frac{1}{N}\sum_{i=1}^{N}\bar{s}_i^{(\theta)}\left(\theta_0,\hat{\alpha}_i\left(\theta_0\right),\hat{\psi}\right)+\frac{1}{N}\sum_{i=1}^{N}\bar{h}_i^{(\theta,\theta)}\left(\bar{\theta},\hat{\alpha}_i\left(\bar{\theta}\right),\hat{\psi}\right)\left(\hat{\theta}-\theta_0\right), \tag{A.14}$$

where, by Lemma A.1(i) and Theorem A.1,

$$\begin{aligned} &\left\|\frac{1}{N}\sum_{i=1}^{N}\bar{h}_i^{(\theta,\theta)}\left(\bar{\theta},\hat{\alpha}_i\left(\bar{\theta}\right),\hat{\psi}\right)-H_0^{(\theta,\theta)}\left(\psi_0\right)\right\| \\ \leq\ &\sup_{\theta,\alpha}\left\|\bar{h}^{(\theta,\theta)}\left(\theta,\alpha,\hat{\psi}\right)-H^{(\theta,\theta)}\left(\theta,\alpha,\psi_0\right)\right\|_\infty+\left\|\frac{1}{N}\sum_{i=1}^{N}H_i^{(\theta,\theta)}\left(\bar{\theta},\hat{\alpha}_i\left(\bar{\theta}\right),\psi_0\right)-H_0^{(\theta,\theta)}\left(\psi_0\right)\right\| \\ =\ &o_P\left(1\right)+o_P\left(1\right). \end{aligned}$$

Next, we expand the first term in eq. (A.14) w.r.t. $\hat{\alpha}_i\left(\theta_0\right)$, $i=1,...,N$,

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}\bar{s}_i^{(\theta)}\left(\theta_0,\hat{\alpha}_i\left(\theta_0\right),\hat{\psi}\right) &= \frac{1}{N}\sum_{i=1}^{N}\bar{s}_{0,i}^{(\theta)}\left(\hat{\psi}\right)+\frac{1}{N}\sum_{i=1}^{N}\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)\left(\hat{\alpha}_i-\alpha_{0,i}\right) \\ &+ \frac{1}{2}\sum_{k=1}^{d_\alpha}\frac{1}{N}\sum_{i=1}^{N}\bar{g}_{0,i}^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)\left(\hat{\alpha}_i-\alpha_{0,i}\right)\left(\hat{\alpha}_{i,k}-\alpha_{0,i,k}\right)+R_N^{(1)}, \end{aligned} \tag{A.15}$$

where, with $\bar{\alpha}_{0,i}$ situated on the line segment connecting $\hat{\bar{\alpha}}_i\left(\theta_0\right)$ to $\alpha_{0,i}\left(\theta_0\right)$ and applying Lemma A.1 and Theorem A.1,

$$\begin{aligned} \left\|R_N^{(1)}\right\| &\leq \frac{1}{2}\sum_{k=1}^{d_\alpha}\max_{i=1,...,N}\left\|\left\{\bar{g}^{(\theta,\alpha,\alpha_k)}\left(\theta_0,\bar{\alpha}\left(\theta_0\right),\hat{\psi}\right)-\bar{g}_0^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)\right\}\right\|_\infty\left\|\hat{\alpha}-\alpha_0\right\|_\infty^2 \\ &\leq \frac{d_\alpha}{2}\left\|\bar{B}_g\right\|_\infty\left\|\hat{\alpha}-\alpha_0\right\|_\infty^3=O_P\left(T^{-3\rho}\right). \end{aligned}$$

For the first-order term on the right-hand side of eq. (A.15), use (3.5) to write

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)(\hat{\alpha}_i-\alpha_{0,i}) \\
=\;&\frac{1}{N}\sum_{i=1}^{N}\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right)+\frac{1}{N}\sum_{i=1}^{N}\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)\bar{v}_i\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right)+R_N^{(2)} \\
=\;&\frac{1}{N}\sum_{i=1}^{N}H_{0,i}^{(\theta,\alpha)}(\psi_0)\,\bar{u}_i\left(\hat{\psi}\right)+\frac{1}{N}\sum_{i=1}^{N}\left\{\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)-H_{0,i}^{(\theta,\alpha)}(\psi_0)\right\}\bar{u}_i\left(\hat{\psi}\right) \\
&+\frac{1}{N}\sum_{i=1}^{N}H_{0,i}^{(\theta,\alpha)}(\psi_0)\,\bar{v}_i\left(\hat{\psi}\right)\bar{u}_i\left(\hat{\psi}\right)+R_N^{(2)}+R_N^{(3)},
\end{aligned}$$

with $R_N^{(2)}=\frac{1}{N}\sum_{i=1}^{N}\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)r_i^{(\alpha)}$ and $R_N^{(3)}=\frac{1}{N}\sum_{i=1}^{N}\left\{\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)-\bar{h}_{0,i}^{(\theta,\alpha)}(\psi_0)\right\}\bar{v}_i\bar{u}_i$. Theorem A.1 and Lemma A.1 yield

$$\begin{aligned}
\left\|R_N^{(2)}\right\| &\leq \left\|\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)\right\|_\infty\left\|r^{(\alpha)}\right\|_\infty=O_P(1)\,O_P\left(T^{-3\rho}\right)=O_P\left(T^{-3\rho}\right), \\
\left\|R_N^{(3)}\right\| &\leq \left\|\bar{h}_{0,i}^{(\theta,\alpha)}\left(\hat{\psi}\right)-\bar{h}_{0,i}^{(\theta,\alpha)}(\psi_0)\right\|_\infty\|\bar{v}\|_\infty\|\bar{u}\|_\infty=O_P\left(T^{-3\rho}\right).
\end{aligned}$$

Each of the second-order terms in (A.15) satisfies, $k=1,...,d_\alpha$,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\bar{g}_{0,i}^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)(\hat{\alpha}_i-\alpha_{0,i})(\hat{\alpha}_{i,k}-\alpha_{0,i,k}) &= \frac{1}{N}\sum_{i=1}^{N}\bar{g}_{0,i}^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)\bar{u}_i\bar{u}_{i,k}+R_{N,k}^{(4)} \\
&= \frac{1}{N}\sum_{i=1}^{N}G_{0,i}^{(\theta,\alpha,\alpha_k)}(\psi_0)\,\bar{u}_i\bar{u}_{i,k}+R_{N,k}^{(4)}+R_{N,k}^{(5)},
\end{aligned}$$

where, using the same arguments as in the analysis of the first-order term,

$$\begin{aligned}
\left\|R_{N,k}^{(4)}\right\| &\leq \left\|\bar{g}_0^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)\right\|_\infty\left\{2\|\bar{v}\|_\infty\|\bar{u}\|_\infty^2+\left\|r^{(\alpha)}\right\|_\infty\right\}=O_P(1)\left\{O_P\left(T^{-3\rho}\right)+O_P\left(T^{-3\rho}\right)\right\} \\
&= O_P\left(T^{-3\rho}\right),
\end{aligned}$$

$$\left\|R_{N,k}^{(5)}\right\|\leq\left\|\bar{g}_0^{(\theta,\alpha,\alpha_k)}\left(\hat{\psi}\right)-G_0^{(\theta,\alpha,\alpha_k)}(\psi_0)\right\|_\infty\|\bar{u}\|_\infty^2=O_P\left(T^{-\rho}\right)O_P\left(T^{-2\rho}\right)=O_P\left(T^{-3\rho}\right).$$

Collecting terms, the claimed expansion is obtained. □

Finally, we expand the leading bias and variance terms in (A.10) w.r.t. $\hat{\psi}$ around ψ_0 to obtain Theorem 3.2:

Proof of Theorem 3.2. Observe that from the definition of $\Phi_N\left(\hat{\psi}\right)$ in Theorem A.2 together with Lemma A.1 and Assumptions 2-3,

$$\Phi_N\left(\hat{\psi}\right)=\Phi_N(\psi_0)+\nabla\Phi_N\left[\hat{\psi}-\psi_0\right]+\frac{1}{2}\nabla^2\Phi_N\left[\hat{\psi}-\psi_0,\hat{\psi}-\psi_0\right]+O_P\left(T^{-3\rho}\right), \quad \text{(A.16)}$$

where

$$\nabla\Phi_N[d\psi] = H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{1}{N}\sum_{i=1}^{N}\left\{\nabla\bar{s}_{0,i}^{(\theta)}[d\psi] + H_{0,i}^{(\theta,\alpha)}(\psi_0)\nabla\bar{u}_i[d\psi]\right\}, \tag{A.17}$$

$$\nabla^2\Phi_N[d\psi,d\psi] = H_0^{(\theta,\theta)}(\psi_0)^{-1}\frac{1}{N}\sum_{i=1}^{N}\left\{\nabla^2\bar{s}_{0,i}^{(\theta)}[d\psi,d\psi] + H_{0,i}^{(\theta,\alpha)}(\psi_0)\nabla^2\bar{u}_i[d\psi,d\psi]\right\}, \tag{A.18}$$

and

$$\nabla\bar{u}_i[d\psi] = -H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\nabla\bar{s}_{0,i}^{(\alpha)}[d\psi],\ \nabla^2\bar{u}_i[d\psi,d\psi] = -H_{0,i}^{(\alpha,\alpha)}(\psi_0)^{-1}\nabla^2\bar{s}_{0,i}^{(\alpha)}[d\psi,d\psi]. \tag{A.19}$$

Here, $\nabla\bar{s}_{0,i}^{(\alpha)}[d\psi] = \sum_{t=1}^{T}\nabla s_{0,i,t}^{(\alpha)}[d\psi]/T$ and similar for other average differentials. We do not need to expand the bias term, on the other hand, since, using that the funtions entering $B_N\left(\hat{\psi}\right)$ are Lipschitz w.r.t. $\hat{\psi}$,

$$\left\|B_N\left(\hat{\psi}\right)/T - B_N(\psi_0)/T\right\| = O_P\left(\left\|\hat{\psi}-\psi_0\right\|/T\right) = O_P\left(T^{-1-\rho}\right).$$

where the remainder term is negiglible. Under Assumptions 2 and 3, eq. (A.16) holds with

$$\begin{aligned}\left\|\nabla\Phi_N\left[\hat{\psi}-\psi_0\right]\right\| &= O_P\left(\left\|\hat{\psi}-\psi_0\right\|\right) = O_P\left(T^{-\rho}\right),\\ \nabla^2\Phi_N\left[\hat{\psi}-\psi_0,\hat{\psi}-\psi_0\right] &= O_P\left(\left\|\hat{\psi}-\psi_0\right\|^2\right) = O_P\left(T^{-2\rho}\right).\end{aligned}$$

Finally, applying parts (ii) and (iii) of Lemma A.2 yield (3.7) and (3.8), respectively, wherewith

$$\Omega(\psi_0) = Var_\infty\left(s_0^{(\theta)}(\psi_0) + E\left[h_0^{(\theta,\alpha)}(\psi_0)\right]u(\psi_0)\right), \tag{A.20}$$

$$\begin{aligned}B(\psi_0) = {} & H_0^{(\theta,\theta)}(\psi_0)^{-1}\left\{Cov_\infty\left(h_0^{(\theta,\alpha)}(\psi_0),u(\psi_0)\right) + Cov_\infty\left(E\left[h_0^{(\theta,\alpha)}(\psi_0)\right]v(\psi_0),u(\psi_0)\right)\right\}\\ & + \frac{1}{2}H_0^{(\theta,\theta)}(\psi_0)^{-1}\sum_{k=1}^{d_\alpha}Cov_\infty\left(E\left[g_0^{(\theta,\alpha,\alpha_k)}(\psi_0)\right]u(\psi_0),u(\psi_0)\right).\end{aligned} \tag{A.21}$$

□

A.3 Remaining proofs

The following theorem is a generalization of Theorem 3.1 in Dhaene and Jochmans (2015):

Theorem A.3. *Suppose that a given estimator $\hat{\theta}$ based on $\{z_{i,t} : t = 1,....,T, i = 1,...,N\}$ satisfies, as $N,T\to\infty$,*

$$\sqrt{NT}\left\{\hat{\theta}-\theta_0-\mathcal{B}_{N,T}\right\} = \sqrt{NT}\Phi_N(\psi_0) + o_P(1) \to^D N(0,N(0,V)),$$

where $\Phi_N(\psi_0) = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \phi(z_{i,t})$ *and* $\mathcal{B}_N = \mathcal{B}_{N,1}/T + \mathcal{B}_{N,2}$ *with* $\mathcal{B}_{N,1} \to^P \mathcal{B}_1$ *and* $\mathcal{B}_{N,2} = \frac{1}{T}\sum_{t=1}^T b_{N,2,t}$ *for some possibly* N*-dependent sequence* $b_{N,2,t}$*. Moreover,* $\{z_{i,t}\}$ *satisfies the assumptions of HK. Then the Jackknife estimator in eq. (2.9) satisfies, if* $\sqrt{NT}/T = O(1)$*,* $\sqrt{NT}\left\{\hat{\theta} - \theta_0\right\} \to^D N(0, N(0, V))$*.*

Combining Theorems A.2 and A.3, we obtain Corollary 3.1.

Proof of Theorem A.3. With

$$\Phi_N^{(1)}(\psi_0) = \frac{2}{NT}\sum_{i=1}^N\sum_{t=1}^{T/2}\phi(z_{i,t}),\ \Phi_N^{(2)}(\psi_0) = \frac{2}{NT}\sum_{i=1}^N\sum_{t=T/2+1}^{T}\phi(z_{i,t})$$

and $\mathcal{B}_N^{(i)} = 2\mathcal{B}_{N,1}^{(i)}/T + \mathcal{B}_{N,2}^{(i)}$, $i = 1, 2$, where

$$\mathcal{B}_{N,2}^{(1)} = \frac{2}{T}\sum_{t=1}^{T/2} b_{N,2,t},\ \mathcal{B}_{N,2}^{(2)} = \frac{2}{T}\sum_{t=T/2+1}^{T} b_{N,2,t},$$

the following hold by the CLT in Lemma A.2:

$$\begin{aligned}\Delta_N &:= \sqrt{NT}\begin{pmatrix}\hat{\theta} - \theta_0 - \mathcal{B}_N\\ \hat{\theta}^{(1)} - \theta_0 - \mathcal{B}_N^{(1)}\\ \hat{\theta}^{(2)} - \theta_0 - \mathcal{B}_N^{(2)}\end{pmatrix} = \sqrt{NT}\begin{pmatrix}\Phi_N(\psi_0)\\ \Phi_N^{(1)}(\psi_0)\\ \Phi_N^{(2)}(\psi_0)\end{pmatrix} + o_P(1)\\ &\to^D N\left(0, \begin{pmatrix}V & V & V\\ V & 2V & 0\\ V & 0 & 2V\end{pmatrix}\right).\end{aligned}$$

Thus, using that $2\mathcal{B}_N - \frac{1}{2}\left(\mathcal{B}_N^{(1)} + \mathcal{B}_N^{(2)}\right) = o_P(1/T)$,

$$\begin{aligned}\sqrt{NT}\left\{\hat{\theta} - \theta_0\right\} &= \sqrt{NT}\left\{2\hat{\theta} - \frac{1}{2}\left(\hat{\theta}_1 + \hat{\theta}_2\right) - \theta_0\right\}\\ &= \sqrt{NT}\left\{2\left(\hat{\theta} - \theta_0 - \mathcal{B}_N\right) - \frac{1}{2}\left(\hat{\theta}_1 - \theta_0 - \mathcal{B}_{N,1} + \hat{\theta}_1 - \theta_0 - \mathcal{B}_{N,2}\right)\right\}\\ &\quad + o_P\left(\sqrt{NT}/T\right)\\ &= (2I_{d_\theta}, -I_{d_\theta}/2, -I_{d_\theta}/2)\,\Delta_N + o_P(1) \to^D N(0, V),\end{aligned}$$

where we have used that

$$(2I_{d_\theta}, -I_{d_\theta}/2, -I_{d_\theta}/2)\begin{pmatrix}V & V & V\\ V & 2V & 0\\ V & 0 & 2V\end{pmatrix}\begin{pmatrix}2I_{d_\theta}\\ -I_{d_\theta}/2\\ -I_{d_\theta}/2\end{pmatrix} = V.$$

□

B Tables and Figures

B.1 Jack-knife

Table 9: Example 1. Homogeneous, θ. $N = 1000, T = 10$.

	Baseline			Half-panel Jackknife		
	Bias	MC std.	RMSE	Bias	MC std.	RMSE
FE	0.204	0.022	0.205	-0.814	0.137	0.826
AFE, Linear interpolation						
$J = 2$	0.027	0.006	0.028	-0.171	0.019	0.172
$J = 3$	0.134	0.017	0.135	-0.286	0.040	0.289
$J = 4$	0.159	0.017	0.160	-0.535	0.069	0.540
$J = 5$	0.185	0.019	0.186	-0.623	0.104	0.632
$J = 10$	0.200	0.021	0.201	-0.778	0.135	0.790
$J = 25$	0.203	0.022	0.205	-0.827	0.144	0.839
$J = 50$	0.204	0.022	0.205	-0.827	0.144	0.840
$J = 100$	0.204	0.022	0.205	-0.824	0.144	0.836

Notes: Shows Monte Carlo results for Example 1 for the homogeneous parameter, θ. We have used 500 Monte Carlo runs and the parameters in Table 4.

Table 10: Example 1. Homogeneous, θ. $N = 1000, T = 30$.

	Baseline			Half-panel Jackknife		
	Bias	MC std.	RMSE	Bias	MC std.	RMSE
FE	0.040	0.004	0.040	-0.023	0.005	0.023
AFE, Linear interpolation						
$J = 2$	-0.028	0.002	0.028	-0.051	0.002	0.051
$J = 3$	0.013	0.003	0.013	-0.032	0.003	0.032
$J = 4$	0.032	0.004	0.032	-0.018	0.004	0.019
$J = 5$	0.032	0.004	0.033	-0.028	0.005	0.028
$J = 10$	0.038	0.004	0.039	-0.023	0.005	0.024
$J = 25$	0.039	0.004	0.040	-0.023	0.005	0.023
$J = 50$	0.039	0.004	0.040	-0.023	0.005	0.023
$J = 100$	0.040	0.004	0.040	-0.023	0.005	0.023

Notes: Shows Monte Carlo results for Example 1 for the homogeneous parameter, θ. We have used 500 Monte Carlo runs and the parameters in Table 4.

C Data

C.1 Income Definitions

In the Danish income registers, we have the following income variables:

$$\begin{aligned}
\underbrace{\text{DISPON_NY}}_{\text{disposable income}} &= \text{SAMLINK_NY} - \underbrace{\text{SKATMVIALT_NY}}_{\text{taxes}} \\
&\quad - \underbrace{\text{QRENTUD2}}_{\text{interest payments}} - \underbrace{\text{UNDERHOL} + \text{TBKONTHJ}}_{\text{alimony+returned benefits}} \\
\underbrace{\text{SAMLINK_NY}}_{\text{total income}} &= \text{PERINDKIALT} + \underbrace{\text{OVSKEJD02_NY} + \text{OVERSKEJD07}}_{\text{imputed rental value}} \\
\underbrace{\text{PERINDKIALT}}_{\text{total monetary income}} &= \underbrace{\text{RENTEINDK}}_{\text{interest income}} + \underbrace{\text{PEROEVRIGFORMUE}}_{\text{other property income}} + \\
&\quad \underbrace{\text{ERHVERVSINDK(_GL)}}_{\text{wages and profits}} + \underbrace{\text{OVERFORSINDK}}_{\text{public transfers}} \\
&\quad + \underbrace{\text{RESUINK(_GL)}}_{\text{other income}}
\end{aligned}$$

We define nominal income for couple i in year t as

$$\begin{aligned}
Y_{it}^{nom} \equiv\ & \text{PERINDKIALT} - \text{RENTEINDK-PEROEVRIGFORMUE} \\
& - \text{SKATMVIALT_NY} - \text{UNDERHOL} - \text{TBKONTHJ}
\end{aligned}$$

C.2 Data Construction

We construct our variables as follows:

1. **Couples** are constructed using *EFALLE* (from BEF).
2. **Birthyear** and **gender** is based on FOED_DAG and KOEN (from BEF). Couple age is the age of the male.
3. **Wealth** A_{it}^{nom} is the total net wealth excluding pensions (FORM from INDH) adjusted upwards with 10 percent of the value of any owned properties (KOEJD from INDH).
4. **Self-Employment** is coded as PSTILL≤ 20 (from IDAP).
5. **Retirement** is coded as PSTILL in $\{50, 55, 92, 93, 94\}$ (from IDAP).
6. **Student** is coded as PSTILL $= 91$ (from IDAP).
7. A couple is coded as **high-skilled** if at least one of them has ≥ 180 months of education (using HFPRIA from UDDA); otherwise it is coded as **low-skilled**.

We additionally calculate nominal cash-on-hand and imputed consumption as

$$M_{it}^{nom} \equiv R \cdot A_{i,t-1}^{nom} + Y_{it}^{nom} \tag{C.1}$$

$$C_{it}^{nom} \equiv M_{it}^{nom} - A_{it}^{nom} \tag{C.2}$$

All variables are subsequently deflated with the consumer price index.

C.3 Sample Selection

We use the following iterative sample selection criteria:

1. Our baseline sample is all couples in the period 1987 and 1996 (both included).
2. Both partners are between age 25 and 59 (both included).
3. The age difference is not larger than 5 years.
4. All observations before and when one is a student is dropped.
5. All observation after and when one is retired is dropped.
6. Neither of them are ever self-employed.
7. Neither of them are ever out of the labor market.
8. Education information is not missing for both partners.
9. We remove all households with fewer than 5 observations satisfying:
   (a) $\frac{M_{it}}{Y_{it}}$, $\frac{C_{it}}{Y_{it}}$, and Y_{it} are not below the 1st percentile or above the 99th percentile by age-year bins.
   (b) $m_{it} \equiv \frac{M_{it}}{P_{it}} \geq -\lambda$
   (c) $a_{it} \equiv \frac{A_{it}}{P_{it}} \geq -\lambda$
   (d) $c_{it} \equiv \frac{C_{it}}{P_{it}} < 0.3$

   Additionally, we do not use consumption for any of the periods where the above requirements are not satisfied.

Table 11 shows how the sample size is affected by these choices.

Table 11: Sample Selection

	Unique Couples	Observations
1. Baseline	1,933,846	12,852,936
2. Age between 25 and 59	1,460,232	9,127,239
3. Age difference ≤ 5 years	1,102,895	7,209,223
4. Not student	1,078,456	7,002,862
5. Not retired	1,040,114	6,612,273
6. Never self-employed	837,050	5,114,581
7. Never out of the labor market	659,347	4,103,660
8. Some education information	656,522	4,093,567
9. More that 5 observations	406,973	3,558,403
10. More that 5 observations - excluding extremes	261,725	1,966,741

Figure 8: Income Growth Factors, G_t.