IFS

Institute for Fiscal Studies

Jack Britton
Ben Waltmann

21/13 Working paper

Revisiting the solution of dynamic discrete choice models: time to bring back Keane and Wolpin (1994)?

UKRI Economic and Social Research Council

Revisiting the solution of dynamic discrete choice models: time to bring back Keane and Wolpin (1994)?[*]

JACK BRITTON
IFS

BEN WALTMANN
Oxford and IFS

May 2021

Abstract

The 'curse of dimensionality' is a common problem in the estimation of dynamic models: as models get more complex, the computational cost of solving these models rises exponentially. Keane and Wolpin (1994) proposed a method for addressing this problem in finite-horizon dynamic discrete choice models by evaluating only a subset of state space points by Monte Carlo integration and interpolating the value of the remainder. This method was widely used in the late 1990s and 2000s but has rarely been used since, as it was found to be unreliable in some settings. In this paper, we develop an improved version of their method that relies on three amendments: systematic sampling, data-guided selection of state space points for Monte Carlo integration, and dispensing with polynomial interpolation when a multicollinearity problem is detected. With these improvements, the Keane and Wolpin (1994) method achieves excellent approximation performance even in a model with a large state space and substantial *ex ante* heterogeneity.

[*]We are grateful to Rob Sauer for many helpful discussions on this topic, and to James Duffy, Zoe Fannon, Mike Keane, and seminar participants at Oxford for useful comments and suggestions. We thank the Centre for the Microeconomic Analysis of Public Policy (ES/T014334/1) for financial support. In addition, Britton acknowledges funding from the British Academy (PF150081). Waltmann acknowledges support from an ESRC doctoral studentship. We also gratefully acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility (DOI code: http://dx.doi.org/10.5281/zenodo.22558).
Corresponding author: Ben Waltmann, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK. Email: ben.waltmann@ifs.org.uk.

1 Introduction

Many economic models are *dynamic* in the sense that current choices affect not only current but also expected future returns. A common problem in the estimation of these models is the 'curse of dimensionality': as models get more complex, the computational cost of solving them rises exponentially with the dimensionality of the state space. Keane and Wolpin (1994) (hereafter KW94) proposed a very general method for addressing this problem for the subclass of dynamic models where agents repeatedly choose from a discrete set of options and the time horizon is finite. Models with this structure are common in labour economics, where they have been used to study career choice and labour supply.

The KW94 method works backwards from the final period. A subset of state space points is evaluated by Monte Carlo integration each period. The value of the remaining points is imputed using a polynomial interpolation method. This saves computation time relative to a full solution, where all state space points are evaluated using Monte Carlo integration. The method was widely used in empirical work in the late 1990s and 2000s.[1] However, it has rarely been used since, as it was found to be unreliable in some settings.[2]

In this paper, we develop an improved version of the KW94 method. Using two versions of the KW94 model as test cases, we show that the improved KW94 method performs very well, both in Monte Carlo simulations at the true parameter values and when the model parameters are estimated from simulated data. Holding the number of integral simulation draws constant, simulated choices of individuals when the improved KW94 approximation is used are nearly as close to 'true' simulated choices as when the full solution is used. When the model is estimated from simulated data, this translates into accurate parameter estimates and precise predictions in a simple policy experiment, again in line with the full solution. These results obtain both in the canoncial KW94 model and in a variation of this model with a larger state space and substantial *ex ante* heterogeneity, where the traditional version of the KW94 method is unreliable.

At the same number of integral simulation draws, our improved KW94 approximation

[1] Important papers in labour economics using the method include Keane and Wolpin (1997), Keane and Wolpin (2001), Imai and Keane (2004), Blau and Gilleskie (2006), Lee and Wolpin (2006), Van der Klaauw and Wolpin (2008), Lee and Wolpin (2010) and Keane and Wolpin (2010). It has also been applied in other fields, including marketing science (Erdem and Keane, 1996), health economics (Crawford and Shum, 2005) and development economics (Todd and Wolpin, 2006).

[2] Nevertheless, there is enduring interest in the KW94 approach: in a recent contribution, Eisenhauer (2019) reproduces the results in KW94 and provides an open-source Python package for implementing the method.

is much faster to compute than a full solution; in the two versions of the KW94 model we look at in this paper, the difference is around an order of magnitude. When we instead hold computation time constant, the improved KW94 method is nearly always more accurate than the full solution except at very high computation times. The reason is that our improved KW94 method focuses computational resources on the most important state space points, which is generally preferable to spreading resources equally. This substantially improves the trade-off between model complexity, accuracy, and computation time for applied researchers.

Our improved version of the KW94 method is based on three amendments to the traditional method. First, we use a systematic sampling method instead of simple random sampling for Monte Carlo integration. Systematic sampling reduces the error in integral simulation for the random subset of state space points that are evaluated by Monte Carlo integration. It also indirectly leads to more accurate interpolation, as the parameters of the interpolation polynomial can be more precisely estimated.

Second, taking a data-guided approach to selecting state space points for evaluation by Monte Carlo integration is preferable to the approach recommended in KW94 of selecting them at random. In particular, drawing state space points for Monte Carlo evaluation from the subset of points that agents reach in the data — with the probability of any particular state space point being evaluated proportional to the number of agents reaching that point — can allow for a degree of accuracy comparable to a full model solution even with a very small number of state space points evaluated. Especially when combined, these two improvements can lead to large reductions in required computation time for a desired level of accuracy.

Third, even with these two amendments, the KW94 method can still be unreliable for models with substantial *ex ante* heterogeneity unless a large number of state space points is evaluated by numerical integration. The reason is that unlike in the canonical KW94 model, the estimation of the interpolation polynomial in these models frequently suffers from a multicollinearity problem. We show that a simple and effective solution to this problem is to dispense with the interpolation step if a substantial multicollinearity problem is detected. Instead, it is better to rely solely on an approximation provided by Jensen's Inequality in these cases.

We conclude that the KW94 method, thus improved, is still a useful tool for applied researchers seeking to estimate finite-horizon discrete choice dynamic programming models with large state spaces. It is especially valuable in cases where the model structure precludes the use of the Conditional Choice Probability (CCP) approach of Hotz and Miller (1993) and

Arcidiacono and Miller (2011). This is commonly the case in labour economics, as the CCP approach generally requires error terms to be additive. Models with non-additive error terms include KW94 and Keane and Wolpin (1997).

The rest of the paper is structured as follows. Section 2 briefly lays out the canonical model of KW94. Section 3 outlines our improvements to the KW94 method and provides Monte Carlo evidence of the gains in approximation performance associated with them in the KW94 model. Section 4 shows that these differences in approximation performance translate into differences in the accuracy of parameter estimates and counterfactual predictions. Section 5 introduces *ex ante* heterogeneity into the KW94 model and investiagates the performance of the different versions of the KW94 method relative to the full solution in that setting. Section 6 concludes.

2 A canonical model

The model of KW94 is a model of occupational choice. In each of $T = 40$ discrete periods of time, agents choose between $K = 4$ different options: blue-collar work, white-collar work, education, and home production. The per-period utility functions are

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \alpha_{11}s_t + \alpha_{12}x_{1t} - \alpha_{13}x_{1t}^2 + \alpha_{14}x_{2t} - \alpha_{15}x_{2t}^2 + \epsilon_{1t}) \tag{1}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \alpha_{21}s_t + \alpha_{22}x_{2t} - \alpha_{23}x_{2t}^2 + \alpha_{24}x_{1t} - \alpha_{25}x_{1t}^2 + \epsilon_{2t}) \tag{2}$$

$$u_{3t} = \beta_0 - \beta_1 I(s_t \geq 13) - \beta_2(1 - d_{3,t-1}) + \epsilon_{3t} \tag{3}$$

$$u_{4t} = \gamma_0 + \epsilon_{4t} \tag{4}$$

w_{1t} and w_{2t} are the agent's (latent) wages in occupation one and two. s_t is the number of periods of schooling accumulated by the beginning of period t. x_{1t} and x_{2t} are an individual's total periods of work experience at the beginning of period t in occupation one and two, respectively. $\boldsymbol{d}_t$ is a vector of indicator variables, where for each element $d_{kt} = 1$ if option k is chosen in period t and $d_{kt} = 0$ otherwise. Hence $d_{3,t-1}$ is an indicator variable of whether schooling was chosen in the previous period. $\boldsymbol{\epsilon}_t \sim N(0, \Sigma)$ is a vector of serially uncorrelated shocks, where Σ is parameterized by the parameter vector $\boldsymbol{a}$. $\boldsymbol{\epsilon}_t$ is known to the agent at the beginning of period t but not before. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{a}\}$ is the full parameter vector of the model.

Individuals have perfect knowledge of the true model and rational expectations about the future. Let $\boldsymbol{S}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}, \epsilon_t\}$ be the vector of state variables, and denote the vector of pre-determined state variable by $\overline{\boldsymbol{S}}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}\}$. The vector of pre-determined state variables evolves deterministically given the previous period's choices: $x_{1,t+1} = x_{1,t} + d_{1,t}$, $x_{2,t+1} = x_{2,t} + d_{2,t}$, $s_{t+1} = s_t + d_{3,t}$, and $d_{3,t-1}$ is simply the previous period's choice. Initial conditions are $x_{11} = x_{21} = 0$ and $s_1 = 10$. All agents are *ex ante* homogenous; different choices arise only due to different draws of ϵ_t.[3]

Then the agent's decision problem in period t can be written recursively as

$$\underset{k \leq K}{\operatorname{argmax}}\, V_{kt}(\boldsymbol{S}_t) \tag{5}$$

where

$$V_{kt}(\boldsymbol{S}_t) = \begin{cases} u_{kt}(\boldsymbol{S}_t) + \delta\, \mathrm{E}\left[\max_{j \leq K} V_{j,t+1}(\boldsymbol{S}_{t+1}) | \overline{\boldsymbol{S}}_t, d_{kt} = 1\right] & \text{if } t < T \\ u_{kt}(\boldsymbol{S}_t) & \text{if } t = T \end{cases} \tag{6}$$

is the *alternative-specific value function*. E denotes the mathematical expectations operator. The discount factor δ is fixed at $\delta = 0.95$.

3 Numerical solution of the canonical model

The agent's problem in this model can be solved by backward recursion using (5) and (6). The key challenge is the evaluation of the conditional expectation in (6), which is often called 'EMAX', as it is an expectation of a maximum. This expectation is taken over the joint distribution of ϵ_t and takes the form:

$$\underbrace{\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]}_{\text{EMAX}} = \int \max_{k \leq K} V_{kt}(\overline{\boldsymbol{S}}_t, \epsilon_t) d\Phi(\epsilon_t) \tag{7}$$

where Φ is the (multivariate normal) cdf of ϵ_t and $\boldsymbol{S}_t$ is written as $(\overline{\boldsymbol{S}}_t, \epsilon_t)$ to emphasize the dependence on ϵ_t.

As there is in general no analytical solution for this integral, it usually has to be evaluated using numerical integration methods at considerable computational cost. It is commonly evaluated using crude Monte Carlo integration. In particular, each expectation is replaced by the

[3] A version of this model that features *ex ante* heterogeneity in the form of different per-period utility functions for different types of agents is introduced in Section 5.

consistent and unbiased estimator:

$$\underbrace{\hat{\mathrm{E}}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]}_{\widehat{EMAX}} = \frac{1}{D}\sum_{d=1}^{D} \max_{k\leq K} V_{kt}(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t^d) \tag{8}$$

where the random vectors $\boldsymbol{\epsilon}_t^d$ are drawn from their assumed distribution in the model and D is the number of simulation draws. This procedure is computationally demanding, as a large number of draws have to be evaluated to ensure reasonable accuracy of the numerical integration procedure, and $163,409$ expectations need to be evaluated in total.[4] While modern computers can perform this operation in seconds, computational speed is still an important concern, as the estimation of model parameters typically requires the model to be solved many thousands of times at different parameter values.

KW94 offer a method for speeding up the solution of models of this type. They propose evaluating only a random subset of expectations in each period by Monte Carlo integration. All other expectations are approximated by

$$\underbrace{\tilde{\mathrm{E}}[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)]}_{\widetilde{\mathrm{EMAX}}} = \underbrace{\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} + g\left(\tilde{V}_{1t}, \ldots, \tilde{V}_{Kt}\right) \tag{9}$$

where 'MAXE' is obtained by exchanging the order of the expectation and max operators, g is an interpolation function, and

$$\tilde{V}_{jt}(\boldsymbol{S}_t) = \underbrace{\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right] \text{ for } j = 1, \ldots, K. \tag{10}$$

As recommended in KW94, it is natural to impose $g(.) \geq 0$, since by Jensen's Inequality, $\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]$ ('MAXE') is a lower bound for $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX').[5]

In KW94's preferred specification, the interpolation function g is parameterized as

$$g\left(\tilde{V}_{1t}(\boldsymbol{S}_t), \ldots, \tilde{V}_{Kt}(\boldsymbol{S}_t)\right) = \pi_0 + \sum_{k=1}^{K} \pi_k \tilde{V}_{kt}(\boldsymbol{S}_t) + \sum_{k=1}^{K} \pi_{K+k}\sqrt{\tilde{V}_{kt}(\boldsymbol{S}_t)}. \tag{11}$$

[4] This number reflects the total number of possible combinations of state variables, added up over 39 periods ($t = 2$ to $t = 40$), and taking into account that the maximum permissible number of years of schooling is $s_{max} = 20$, and that it is not possible to have $d_{3,t-1} = 1$ and $s_t = s_0$ or $d_{3,t-1} = 0$ and $s_t = s_0 + t - 1$.

[5] Intuitively, MAXE represents the value of a 'Plan A', whereas EMAX takes into account the value of 'Plans B to D'. As shown in Appendix B, tighter analytical bounds for EMAX are available, but imposing them leads to minor gains at best.

The parameters of this interpolation function are estimated from the subset of expectations evaluated by Monte Carlo integration in each period of the recursion. They are obtained by OLS regression, with

$$y(\boldsymbol{S}_t) = \underbrace{\hat{\mathrm{E}}[\max_{k \le K} V_{kt}(\boldsymbol{S}_t)]}_{\widehat{EMAX}} - \underbrace{\max_{k \le K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\text{MAXE}}$$

as the dependent variable.[6]

As shown in KW94 and confirmed in section 3.4, this method generally provides a good approximation to the full model solution at much lower computational cost. However, the accuracy of the approximation deteriorates substantially as the number of state space points evaluated by Monte Carlo integration is reduced. The method can also be unreliable in some settings; an important case is the estimation of models that feature substantial *ex ante* heterogeneity.

We propose three amendments to the method that address these issues. The first is to use a systematic sampling method for Monte Carlo integration instead of simple random sampling. The second is to select expectations to be evaluated by Monte Carlo integration not at random but based on the part of the state space that agents actually reach in the data. The third detects and avoids multicollinearity problems in the interpolation regression. Together, these changes dramatically improve the approximation performance of the KW94 method, as we show in the remainder of this paper.[7]

3.1 Improvement 1: systematic sampling

The use of systematic sampling can lead to large improvements in the approximation performance of the KW94 method.[8] Systematic sampling guarantees better coverage of the integrals evaluated using Monte Carlo simulation. As a result, the approximation error is generally $O(D^{-1})$ instead of $O_p(D^{-1/2})$ for simple random sampling, where D is the number of simulaton draws (Geweke, 1996). This increases the efficiency not only of the Monte Carlo simulator

[6] More specifically, $\boldsymbol{\pi} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ where $\boldsymbol{X}$ is the data matrix corresponding to (11), with the number of rows of $\boldsymbol{X}$ equal to the number of state space points evaluated by Monte Carlo integration. This presupposes that $\boldsymbol{X}'\boldsymbol{X}$ is invertible, which is not guaranteed. The case when $\boldsymbol{X}'\boldsymbol{X}$ is (near-)singular is especially relevant in models with *ex ante* heterogeneity, and is discussed in section 3.3. In the canonical model, (near-)singular moment matrices $\boldsymbol{X}'\boldsymbol{X}$ almost never occur outside periods $t = 2$ and $t = 3$ at the true parameters. In those early periods, the number of possible state space points is low enough for all to be evaluated by Monte Carlo integration, obviating the need for interpolation.

[7] Appendix C shows that our improved KW94 method is also preferable to various other methods for approximating EMAX that use the MAXE approximation.

[8] As shown in Appendix D.2, it can also substantially improve the performance of the full solution method.

$\widehat{EMAX}$ given by (8), but also of the interpolation estimator $\widetilde{\mathrm{EMAX}}$ given by (9), as the dependent variable of the interpolation regression will be subject to less error.

The systematic sampling algorithm used in the rest of the paper is:

1. Draw u_{kt} from the standard uniform distribution for each $k \leq K$ and $t \in [2, T]$.

2. Calculate
$$\left(\eta_{kt}^{d}\right)_{d=1}^{D} = \Phi^{-1}\left(\frac{d - u_{kt}}{D}\right)$$
for each $k \leq K$ and $t \in [2, T]$.

3. Randomly permute the elments of each $\left(\eta_{kt}^{d}\right)_{d=1}^{D}$ to obtain $\left(\tilde{\eta}_{kt}^{d}\right)_{d=1}^{D}$

4. Obtain $\boldsymbol{\epsilon}_t^d = \boldsymbol{A}\tilde{\boldsymbol{\eta}}_t^d$, where $\boldsymbol{A}$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}$.

Other variance reduction techniques are likely to have similar positive effects on approximation performance. In particular, as we show in Appendix D, using Halton (1964) draws instead of our systematic sampling scheme even leads to slightly better performance. The advantages of our simple scheme are, first, that it is guaranteed to be unbiased, second, that draws are guaranteed to be uncorrelated across both options and periods, and third, that the algorithm is simple to implement for any number of options and periods. The comparison with Halton draws suggests that the cost of these advantages in terms of approximation performance is small.[9]

3.2 Improvement 2: data-driven selection of points for Monte Carlo integration

Our second amendment to the KW94 method is the data-driven selection of state space points to be evaluated using Monte Carlo integration. The method for selecting nodes that we have found to work best is to select nodes for numerical integration that agents actually reach in the data, with more commonly reached nodes selected with a higher probability. Intuitively, the value of nodes that are commonly reached in the data is likely to be more important than that of nodes that are never actually chosen, some of which may not be "real options" given the payoff structure of the model.

A number of different selection schemes will achieve this basic objective. The particular algorithm used in the rest of the paper is:

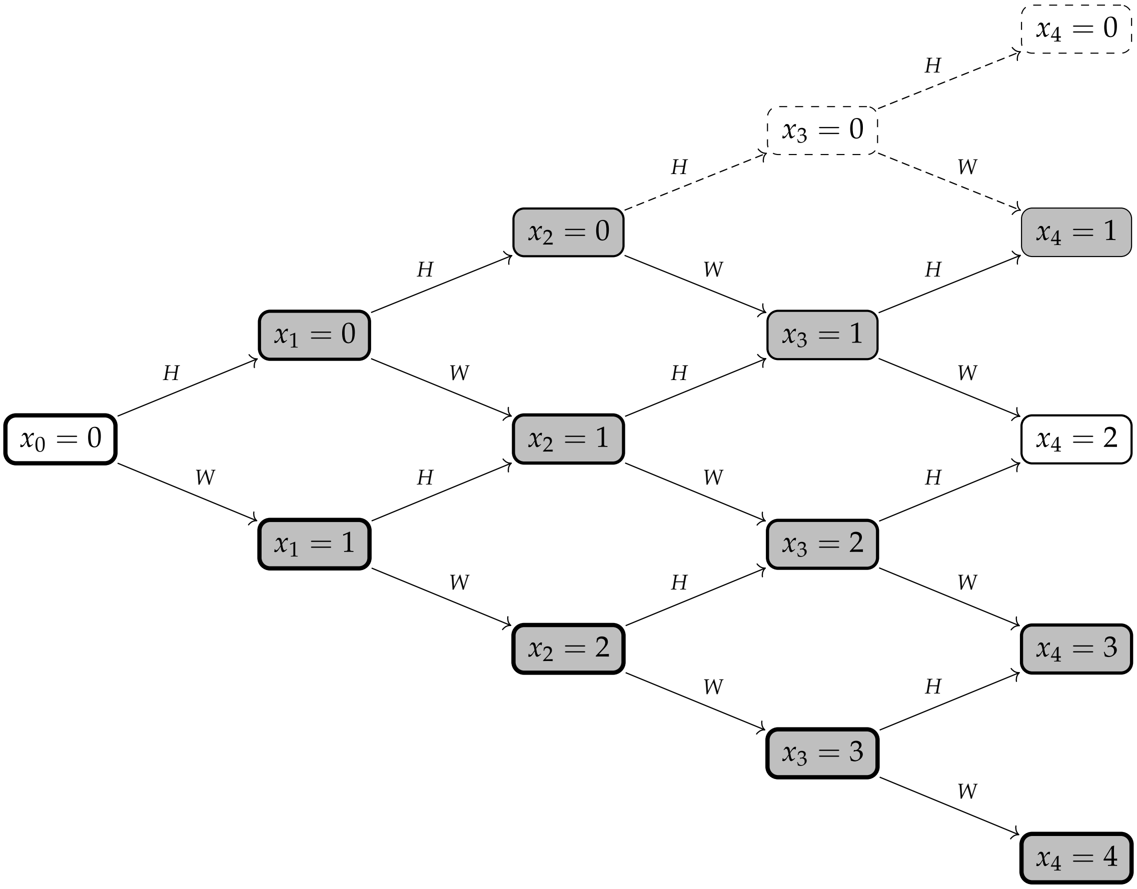
[9] For an intuitive discussion of how systematic sampling and other variance reduction techniques improve on simple random sampling, see Section 9.3 in Train (2009).

1. Permute the data so that individuals appear in random order.

2. In each period, evaluate the nodes reached in the data, working in the order that individuals appear in the data. Continue until all nodes reached in the data have been evaluated, or the maximum number of nodes per period to be evaluated by Monte Carlo integration has been reached.

3. In the former case, draw additional nodes to be evaluated by Monte Carlo integration at random from the remaining set.[10]

Figure 1: State space map for a model with $K = 2$ and $T = 5$



Notes. This graph illustrates the proposed algorithm for selecting state space points for Monte Carlo integation using a minimal example. An explanation is given in the text.

To illustrate this algorithm, consider how it might be applied to a simple dynamic labour force participation model in the spirit of Eckstein and Wolpin (1989) with $T = 5$ periods and a binary choice (so $K = 2$) between market work ('W') and home production ('H'). Suppose

[10] Like the systematic sampling algorithm, this algorithm was primarily chosen for its simplicity; it is unlikely to be the optimal choice.

there is only one state variable that persists across periods: work experience. The stock of work experience is equal to the number of periods spent in market work. The econometrician observes a balanced panel of choices $d_{it} \in \{W, H\}$ for each individual i and time period $t \leq T$.

Figure 1 maps out the state space for this model. There are a total of 15 state space points: in each period t, the stock of work experience up to t can take t different values between $x_t = 0$ (market work in no previous period) and $x_t = t - 1$ (market work in every previous period). Agents can move between state space points as indicated by the arrows. The expected value of being at each state space point can be obtained by backward recursion as in the canonical KW94 model.

Now suppose an approximate solution of this model were to be obtained using the approximation method of KW94, with a maximum of $M = 3$ state space points per period evaluated using Monte Carlo integration, and the rest using an interpolation procedure as outlined. Assume further that in fact most agents work in most periods, so state space points towards the bottom of Figure 1 are reached much more frequently than those near the top, as indicated by the outline thickness. In particular, suppose that all agents in the data have gained at least one year of work experience when they reach period $t = 4$, so the state space points where $x_3 = 0$ and $x_4 = 0$ are never reached (indicated by dashed lines).

Our data-driven selection scheme ensures that nodes that are commonly reached (such as those along the bottom edge of Figure 1) are evaluated with high probability. The intuition behind this procedure is that higher precision is crucial for nodes that are on or near the optimal path for most agents, but less important for those that are not. By evaluating state space points along 'popular' paths, our selection scheme thus achieves a more efficient allocation of computational resources.

The shading in Figure 1 marks the state space points that might be selected for Monte Carlo integration by our procedure in this example. In periods $t = 2$ and $t = 3$, all state space points are evaluated, as $M = 3$ is larger than or equal to the total number of state space points that can be reached.[11] In period $t = 4$, all nodes that are reached in the data will be evaluated. In period $t = 5$, the number of state space points reached in the data is greater than M, so the outcome of the selection algorithm depends on the random ordering of individuals in step 1; one possible outcome is shown.

[11] The expected value of $x_0 = 0$ need not be evaluated, as it is irrelevant for agents' choices.

3.3 Improvement 3: no interpolation when multicollinearity is detected

As the regressors in the interpolation regression (11) are (square roots of) the differences between the option with the highest expected value and the expected value of each option, the value of the regressors relating to the option with the highest expected value is in each case zero. If in a given period one option nearly always has the highest expected value, regardless of an agent's previous path through the model, the interpolation regression will therefore suffer from a multicollinearity problem. This problem is particularly acute in models with substantial *ex ante* heterogeneity, as an agent's type in these models may to a large extent determine their path through the model.

As a result, the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ can be singular or nearly singular. In the former case, the matrix will not be invertible, so it will not be possible to perform the interpolation regression at all. In the latter case, interpolation will be possible, but the parameters will be very sensitive to the values for the small number of nodes where the dominant option does not have the highest expected value.

For the approximation results below, we always set $g(.) = 0$ when the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular. However, this leaves two problems. First, in cases where $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular, the approximation can still go badly awry, resulting in a low proportion of correct choices and very poor fit, as approximation errors in one period affect choices in all earlier periods. Second, dispensing with the interpolation regression when $\boldsymbol{X}'\boldsymbol{X}$ is singular creates discontinuous jumps in the objective surface at the boundary between singularity and near-singularity, potentially impairing estimation. Our third suggestion for applied researchers, therefore, is to set $g(.) = 0$ in cases where $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular as well, and smoothly interpolate between that case and the regular interpolation regression.[12]

One way to accomplish this is to use the (1-norm) *condition number* for inversion of $\boldsymbol{X}'\boldsymbol{X}$, $\kappa(\boldsymbol{X}'\boldsymbol{X})$. We call $\boldsymbol{X}'\boldsymbol{X}$ "nearly singular" if and only if $\kappa(\boldsymbol{X}'\boldsymbol{X}) > 1/\epsilon$, where $\epsilon = 2^{-52} \approx 2.2 \times 10^{-16}$ is the machine epsilon for floating point arithmetic on 64-bit computer systems. For smooth interpolation, we define the polynomial smoothing kernel:

[12] An alternative approach would be to regularize the least squares problem using ridge regression (Hoerl and Kennard, 1970).

$$K(v) = \begin{cases} 0 & \text{for } v < 0 \\ \frac{v^2}{2\phi^2} & \text{for } 0 \leq v < \phi \\ -\frac{v^2}{2\phi^2} + \frac{2v}{\phi} - 1 & \text{for } \phi \leq v < 2\phi \\ 1 & \text{for } v \geq 2\phi \end{cases} \tag{12}$$

where ϕ is a smoothing parameter that we set to $\phi = 50$.[13] We then set $v = \frac{1}{\epsilon\kappa(\boldsymbol{X}'\boldsymbol{X})} - 1$ and replace (9) by

$$\underbrace{\tilde{\mathrm{E}}[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)]}_{\widetilde{\mathrm{EMAX}}} = \underbrace{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} + K(v) g\left(\tilde{V}_{1t}(\boldsymbol{S}_t), \ldots, \tilde{V}_{Kt}(\boldsymbol{S}_t)\right). \tag{13}$$

3.4 Monte Carlo evidence: approximation performance

We evaluate the performance of different approximation methods using Monte Carlo simulations. All simulations are based on the same parameter values as the main results in KW94.[14] We first simulate 500 balanced panels of $N = 10,000$ individuals, recording their choices (and wages, in case a market work option is chosen) in each period. All of these simulations rely on the same 'true' model solution, obtained by fully solving the model by Monte Carlo integration using $D = 80,000$ simple random draws. We then compare these 500 simulated datasets to 500 datasets obtained using different approximation methods.

We use two different metrics to assess the performance of different methods. The first is the share of choices in the simulated datasets based on each approximation method that are the same as in the datasets simulated using the 'true' model solution, holding the sequence of realized shocks $(\epsilon_{it})_{t=1}^{T}$ constant for each individual i. As realized shocks are held constant, any difference in agents' choices will be the result of approximation errors.

Our second metric is the weighted sum of squared deviations from key moments of the simulated data: the share of individuals choosing each option in each period, the mean log wage in each market occupation in each period and the variance of log wages in each occupation in each period.[15] The weight matrix is the inverse of the diagonal of the variance matrix

[13] A similar polynomial kernel is recommended in Hajivassiliou, McFadden and Ruud (1996).

[14] More specifically, these are the parameter values corresponding to their 'data set one'. For individual parameter values, see Appendix E of this paper or Table 1 of KW94.

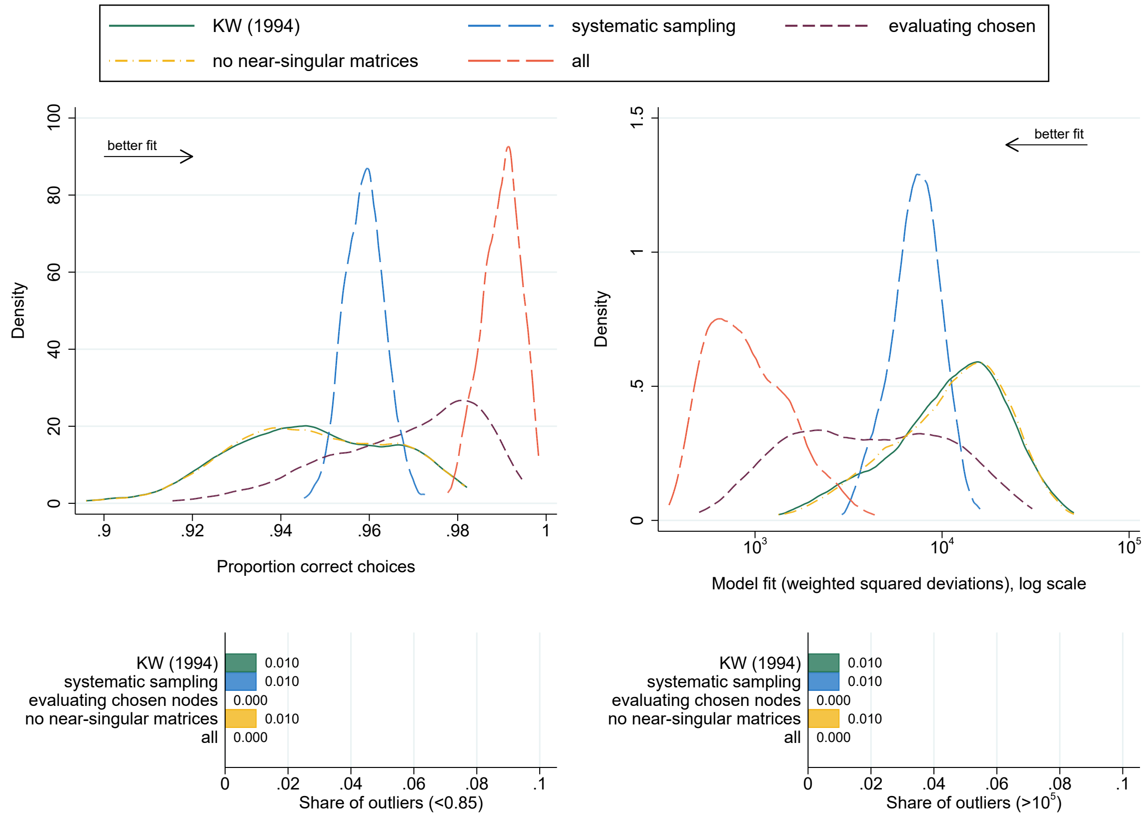
[15] Some choice shares are excluded to avoid linear dependence.

of the vector of data moments. Formally, our measure of model fit φ is given by

$$\varphi(\boldsymbol{\theta}) = (\boldsymbol{m} - \hat{\boldsymbol{m}})'(\text{diag}(\hat{\boldsymbol{V}}))^{-1}(\boldsymbol{m} - \hat{\boldsymbol{m}}) \tag{14}$$

where $\boldsymbol{m}$ is the vector of moments and $\hat{\boldsymbol{V}}$ is the sample variance matrix of $\boldsymbol{m}$.[16]

Figure 2: Approximation performance of different versions of the Keane and Wolpin method



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

For this second metric, not only the approximation draws but also the realized shocks $(\epsilon_{it})_{t=1}^{T}$ are allowed to vary. While this measure is less straightforward to interpret than the share of correct choices, it is a more dependable metric of the distance in simulation outcomes between the true model and different approximations, as it does not rely on knowledge of

[16] This weight matrix is commonly used in empirical work. Examples include Blundell et al. (2016), Morten (2019) and Blundell et al. (2021).

unobservables (the realized shocks $(\epsilon_{it})_{t=1}^{T}$).[17]

Figure 2 shows how our amendments improve the KW94 method when a moderate number of state space points ($M = 500$) is evaluated each period. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two metrics of approximation performance.

The traditional KW94 method performs well on the whole, but the fit varies substantially across approximations run. The share of 'correct' choices varies roughly between 90 percent and 98 percent (disregarding outliers). The model fit measured by weighted squared deviations also varies substantially between approximation runs (more than an order of magnitude), even if outliers are disregarded.

Both the use of systematic sampling and the evaluation of chosen nodes substantially improve performance on their own. Our third improvement has little effect on approximation performance in the KW94 model at the true parameters, as (near-)singular moment matrices $\boldsymbol{X}'\boldsymbol{X}$ typically only occur in periods $t = 2$ and $t = 3$ when the number of possible state space points is low enough for all expectations to be evaluated by Monte Carlo integration. The combination of all amendments is much better still. For the variation combining all changes, the proportion of correct choices is always above 97 percent. The median weighted sum of squared deviations from true moment values is roughly an order of magnitude lower in the variation with all changes than using the traditional KW94 method.

Figure 3 shows how our changes to the KW94 method affect the trade-off between accuracy and computation time. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. Again, the two panels show our two metrics of approximation performance.[18]

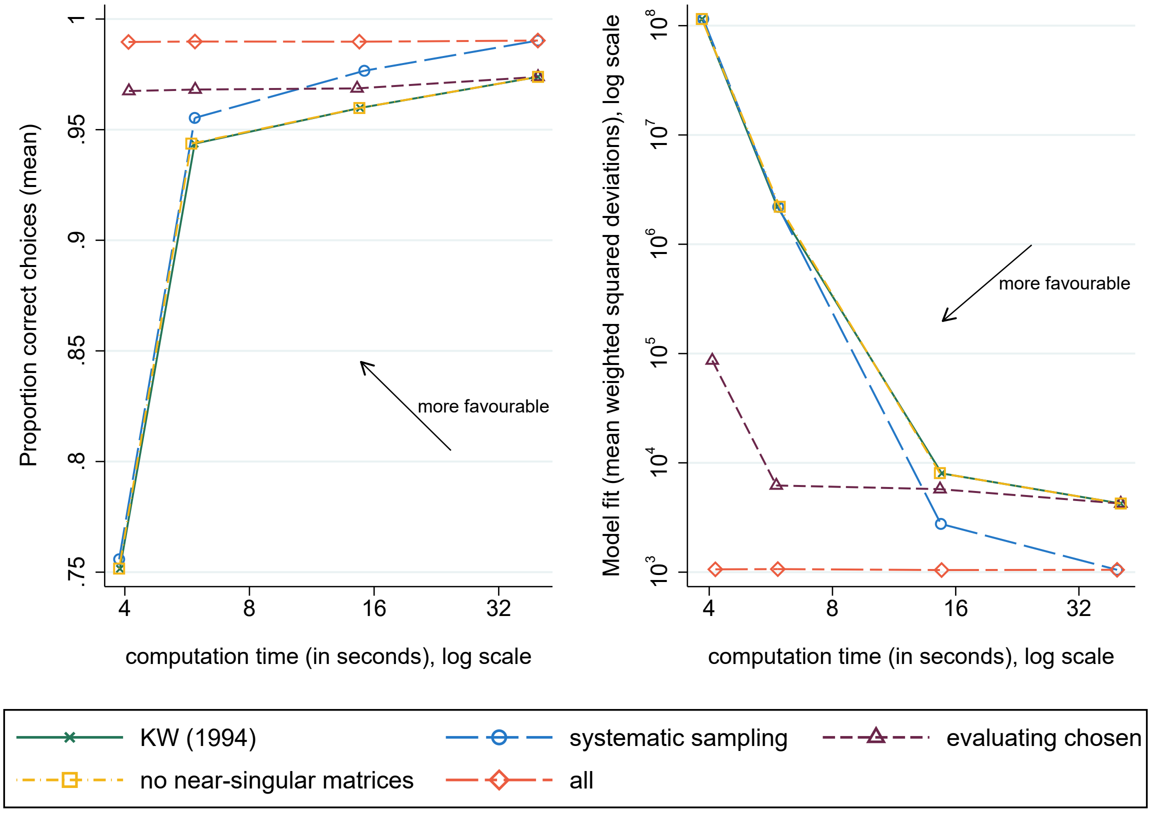
Our improvements fundamentally change the nature of the trade-off between accuracy and computation time for the KW94 method. In the traditional version, the trade-off is quite steep even with a large number of nodes evaluated using Monte Carlo simulation. In contrast, when evaluation nodes are selected systematically, the trade-off essentially disappears for moderate numbers of nodes evaluated by Monte Carlo integration. With all improvements, excellent approximation performance is achieved even when at most $M = 250$ nodes are evaluated each

[17] Importantly, using this second metric, we can assess whether the advantage of evaluating nodes chosen in the data depends on assuming the same realized shocks.

[18] The apparent difference in the shape of the trade-off across the two measures is mainly a result of the higher sensitivity to outliers of the average sum of weighted squared deviations compared to the average share of correct choices.

period.

Figure 3: Trade-off between accuracy and computation time



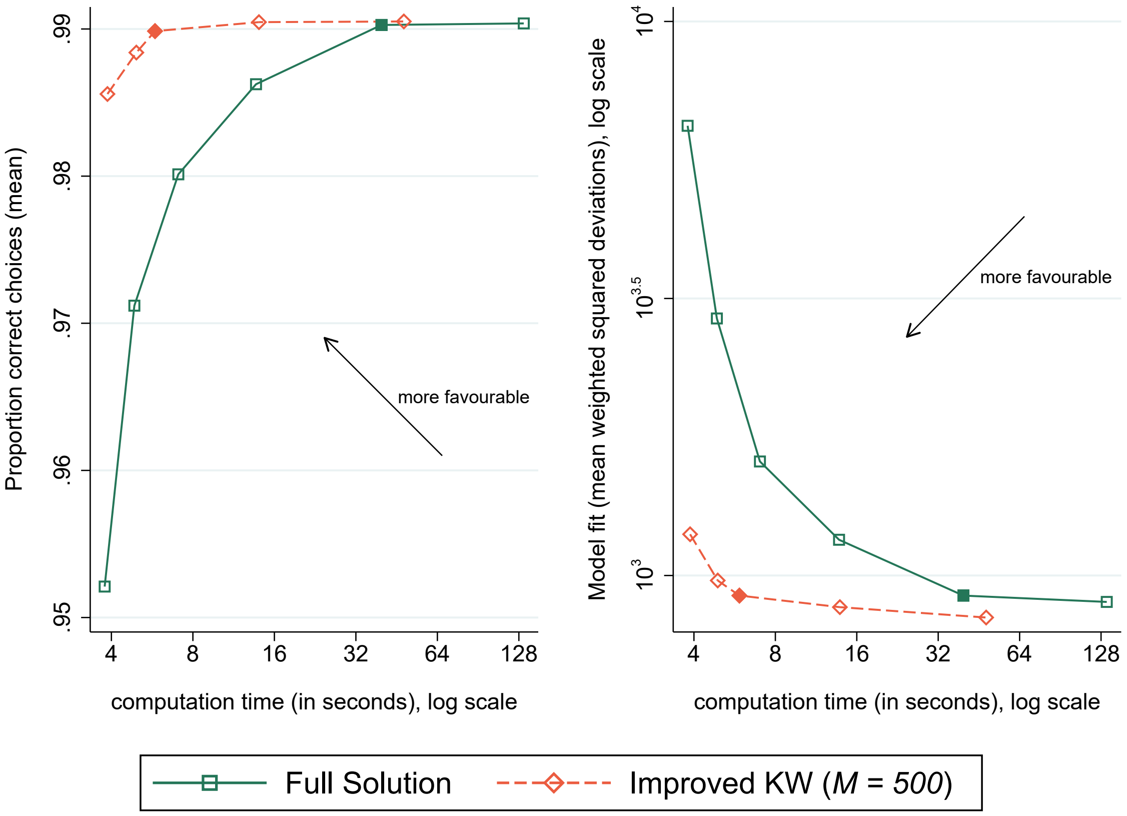
Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Figure 4 compares our improved KW94 method with $M = 500$ evaluation points to the full solution, with the number of draws adjusted to achieve comparable computation times. At nearly all levels of computational expense, the improved KW94 method delivers superior approximation accuracy, with a larger lead at shorter computation times. The full solution reaches similar accuracy only at computation times larger than around one minute.

4 Estimation of the canonical model

This section shows that these gains in approximation performance for a known set of parameters translate into similar gains in estimation performance when the parameters are unknown and need to be estimated from observed data. Using the improved KW94 method, the parameters of the canonical model can be estimated with a high degree of accuracy even when only a small number of nodes are evaluated using Monte Carlo integration. This in turn translates into more accurate predictions in a simple policy experiment.

Figure 4: Trade-off between accuracy and computation time: comparison to full solution



Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $N = 10,000$ individuals' choices were simulated. For the full solution, the six data points shown represent, from left to right, the solution with $D = 50$, $D = 100$, $D = 200$, $D = 500$, $D = 2000$ and $D = 5000$ draws. For the improved KW94 method, the five data points shown represent, from left to right, the solution with $D = 500$, $D = 1000$, $D = 2000$, $D = 5000$, and $D = 20,000$ draws. Filled symbols indicate $D = 2000$, the number of Monte Carlo draws used in other graphs. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel). For comparability, systematic sampling was used for both methods.

These results obtain regardless of whether the model is estimated using Simulated Maximum Likelihood as in the original KW94 contribution and much of the literature at the time, or using a simple Simulated Method of Moments or Indirect Inference method, as has become more popular recently. Simulated Maximum Likelihood can provide more accurate results, but this typically comes at the cost of higher computation time, as numerical integration is required to evaluate the likelihood function.[19] We present Simulated Maximum Likelihood results below; analogous Simulated Method of Moments results are presented in Appendix A.

As in KW94, all of the results in this section are presented for a fixed (maximum) number of nodes evaluated by Monte Carlo integration, a fixed number of numerical integration draws for both solution and estimation, and with the smoothing parameter held constant. The motivation for this is that it makes our method more transparent and the results easier to interpret. In actual estimation, it would be advisable to increase the number of nodes evaluated by Monte Carlo integration, increase the number of integration draws, and reduce the amount of smoothing as the algorithm converges.[20]

In order to avoid spuriously positive results, we start the estimation procedures not at the true values but at a set of values that an empirical researcher without knowledge of the true parameter values might reasonably have picked as starting values. In particular, for the wage parameters, we use the parameters of two separate bivariate sample selection models estimated from the data.[21] We further set the value of home production γ_0 to one standard deviation below the median blue collar wage, and the standard deviation of the non-wage options to the average estimated standard deviation of the market-work options. All other parameters are set to zero.

Our Simulated Maximum Likelihood procedure closely follows KW94. The log-likelihood function takes the form

$$l(\theta) = \sum_{i=1}^{N} \left[\log f(\boldsymbol{d}_{i1}, w_{i1}) + \sum_{t=2}^{T} \log f(\boldsymbol{d}_{it}, w_{it} | \{\boldsymbol{d}_{i,s}\}_{s=1}^{s=t-1}) \right]. \tag{15}$$

where $f(\boldsymbol{d}_{it}, w_{it})$ is the joint density of agent i's choice in period t and the corresponding observed wage. If no wage is observed, because the agent chose schooling ($k = 3$) or home production ($k = 4$), $f(\boldsymbol{d}_{it}, w_{it})$ is simply the probability of the observed choice $P(\boldsymbol{d}_{it})$ conditional

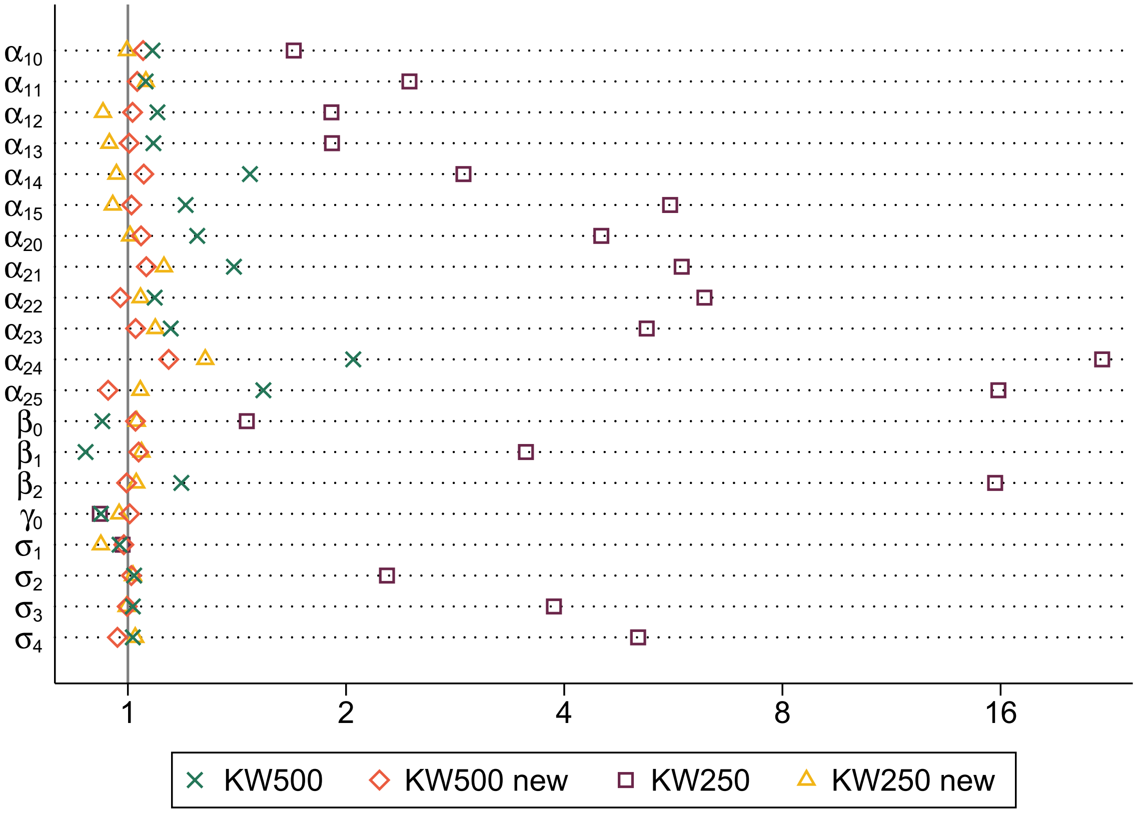
[19] See Eisenhauer, Heckman and Mosso (2015) for a comparison of the two approaches in the context of dynamic discrete choice models.

[20] One way of implementing this would be as a two-step procedure, where (at least) one final step of a (Quasi-)Newton algorithm would be performed using the full solution, a higher number of draws, and less smoothing. This would likely yield additional accuracy gains at little computational cost. For theoretical arguments in favour of this two-step approach, see section 3.1 of Hajivassiliou (2000) and Kristensen and Salanié (2017). Kristensen and Salanié (2017) and Bruins et al. (2018) present Monte Carlo evidence suggesting that this approach also works well in pratice.

[21] Two natural exclusion restrictions arise from the full model: it is known that which period an individual is in and whether school was chosen in the previous period can both affect continuation values and thus choices but not wages. We estimated the selection models using Maximum Likelihood (assuming normally distributed errors). Using the Heckman (1979) two-step method would have resulted in starting values at a similar distance from the the true parameters.

on past choices.

Figure 5: Root mean square error of parameter estimates relative to full solution



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 24 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated datasets; the 6 estimated parameter sets (20%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated dataset used for estimation was a balanced panel of $N = 2,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 15 steps of the BHHH algorithm were performed. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

We evaluate the density $f(\boldsymbol{d}_{it}, w_{it})$ using Monte Carlo integration. The probabilities that enter the likelihood function are smoothed using the smoothed-logit simulator first proposed by McFadden (1989).[22] This is necessary to avoid numerical problems related to simulated probabilities of zero; it has the important added advantage that it leads to a smooth log-likelihood function and thus permits the use of fast and easily parallelized gradient-based optimization routines.[23]

[22] Like KW94, we use a tuning parameter of 500.

[23] We use the BHHH algorithm (Berndt et al., 1974), as extensive experimentation has shown this algorithm to be much more efficient than other commonly used algorithms such as Simplex, Simulated Annealing, or BFGS.

Estimation performance. Figure 5 shows the root mean squared error of all estimated model parameters, relative to the root mean squared error when the full solution with the same number of draws is used. Thirty estimation runs were performed with different simulation draws for the model solution, but with the likelihood simulation draws held constant in order to minimize statistical noise unrelated to the different approximation methods. The six estimation runs (20 percent) with the lowest simulated likelihood at the estimated parameters were dropped to guard against outliers and numerical problems.[24] In each case, the parameters were estimated from a balanced panel of $N = 2000$ people that was simulated from the 'true' model. $D = 2000$ solution draws were used in model solution and each density $f(\boldsymbol{d}_{it}, w_{it})$ was simulated using 200 Halton draws.[25] For each estimation run, 15 steps of the BHHH algorithm were performed; further iterations were found not to change estimated parameter values substantially regardless of the solution method. The root mean squared error was chosen as a measure of accuracy as it captures both bias in the estimated parameter values and their variability.[26]

The root mean squared errors for the improved version of the KW94 method are very close to the root mean squared errors for the full solution, even when at most $M = 250$ state space points are evaluated per period. Root mean squared errors are mostly slightly higher for the traditional version of the KW94 method with $M = 500$ and much higher with $M = 250$.

Policy experiment. In order to gauge the economic significance of these differences in estimation performance, we follow KW94 and perform a simple policy experiment: we investigate the impact of a \$500 tuition subsidy on time spent in each occupation. We simulate two balanced panels from the model for each of the 24 sets of estimated parameters (with and without the subsidy). Each dataset contained $N = 10,000$ individuals, and $D = 2,000$ draws were used in Monte Carlo Integration. We used new draws for both realized shocks and numerical integration. In all cases, simulations were performed using the full solution method.

[24] Estimation failures, marked by a numerical derivative of zero or estimated parameter values implausibly far away from the starting values, were assigned a likelihood of negative infinity (the number of such failures was never higher than six, the number of dropped estimation runs).

[25] The same draws were used for all individuals, and only varied across period t and option k. Halton draws were used to maximize accuracy at a given computational cost.

[26] The *relative* rather than *absolute* root mean squared error is shown, as the parameters have very differents scales. For separate results for bias and standard deviation (comparable with Keane and Wolpin, 1994), see Appendix E.

Table 1: $500 tuition fee subsidy: Simulated Maximum Likelihood

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-3.34	-.126	-3.621	-2.31	-3.596	-3.575	-3.595
	(.119)	(.010)	(.193)	(1.652)	(.213)	(.359)	(.180)
White Collar	2.079	-.235	2.282	1.387	2.257	2.376	2.256
	(.109)	(.013)	(.180)	(1.291)	(.194)	(.402)	(.170)
School	1.461	.408	1.535	1.061	1.532	1.399	1.533
	(.026)	(.005)	(.042)	(.456)	(.045)	(.043)	(.037)
Home	-.199	-.048	-.195	-.138	-.192	-.199	-.194
	(.011)	(.003)	(.015)	(.082)	(.017)	(.032)	(.014)

Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Two balanced panels were simulated from the model for each of 24 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each dataset contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration were different from those used in estimation. In all cases, simulations were performed using the full solution method.

Table 1 shows the estimated effect of the subsidy on the average time spent in each occupation. At the true parameters, the subsidy leads to a rise in the average number of years individuals spend in education and white-collar work, and a fall in the number of years spent in blue-collar work and home production. The column labelled 'Starting' shows the simulated effect when the starting parameters are used instead; even though these parameters are in the vicinity of the true parameters, the simulated effect is very different compared to the simulated effect when the true parameters are used, underlining the importance of accurately estimated parameters.

Simulation based on parameters estimated using the full solution comes close to simulation results obtained using the true parameters, but the maginitude of the effect is slightly overestimated on average. Using the traditional KW94 approximation method with $M = 500$ evaluated nodes and the improved method with $M = 250$ and $M = 500$ nodes, similar simulation values are obtained on average. However, crucially, the improved method leads to much lower standard deviations of the simulated effect. The difference is especially dramatic when at most $M = 250$ nodes are evaluated by Monte Carlo integration each period: in that case, the improved method still comes close to the results using the full solution, whereas the traditional method yields very inaccurate results, echoing the findings for the root mean squared error of the parameter estimates.

5 A model with unobserved heterogeneity

This section introduces unobserved *ex ante* heterogeneity across individuals into the KW94 model and evaluates the performance of both the traditional KW94 approximation method and our improved version in that setting. The motivation is that such heterogeneity is a standard assumption of nearly all applied work in labour economics that uses dynamic discrete choice models.[27] As shown below, the extent of *ex ante* heterogeneity can have a large impact on the performance of different approximation methods. In particular, the traditional method of KW94 can be unreliable when *ex ante* heterogeneity is large. This has likely been a key reason for the recent decline in the use of this method among applied researchers.

We modify the canonical model of KW94 to allow for two different types of agent, where the type is known to the agent but not to the econometrician. An agent's type influences their payoffs from market work as follows:

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \zeta_1 1(\tau_i = 1) + \alpha_{11} s_t + \alpha_{12} x_{1t} - \alpha_{13} x_{1t}^2 + \alpha_{14} x_{2t} - \alpha_{15} x_{2t}^2 + \epsilon_{1t}) \tag{16}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \zeta_2 1(\tau_i = 2) + \alpha_{21} s_t + \alpha_{22} x_{2t} - \alpha_{23} x_{2t}^2 + \alpha_{24} x_{1t} - \alpha_{25} x_{1t}^2 + \epsilon_{2t}) \tag{17}$$

where τ_i is agent i's type and $1(.)$ is the indicator function. We set $\zeta_1 = 0.05$ and $\zeta_2 = 0.1$; all other parameters are unchanged. This implies that type 1 agents enjoy a wage premium of roughly 5% in blue-collar work, and type 2 agents enjoy a wage premium of around 10% in white-collar work. The share of each type is held fixed at one half.[28]

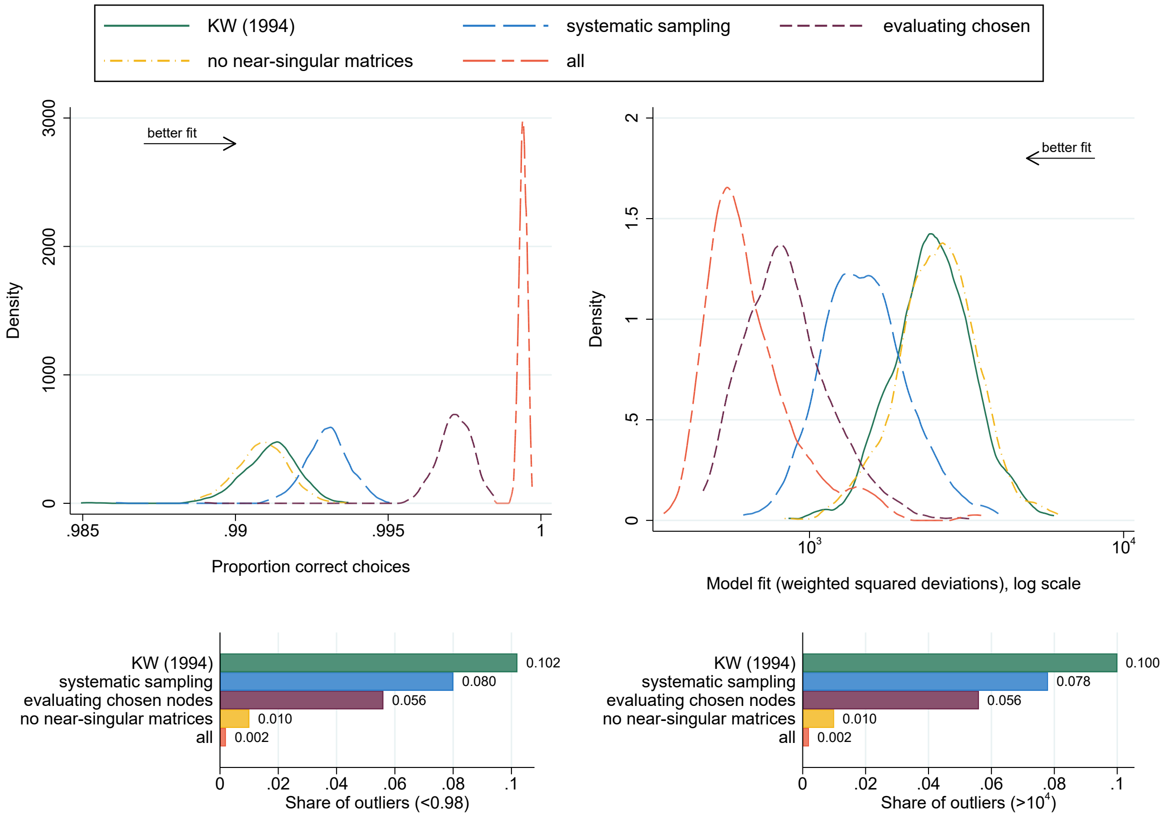
5.1 Approximate solution

One difference arising from the introduction of unobserved heterogeneity into the model is that the state space is now twice as large, as the model now needs to be solved separately for each type. This roughly doubles the computation time required for solving the model regardless of which method is used. As discussed in section 3.3, a particular challenge for the KW94 method in models with *ex ante* heterogeneity is that multicollinearity problems in the interpolation regression are especially likely to occur.

[27] Notably, this includes Keane and Wolpin (1997), which first applied the KW94 method in empirical work.

[28] We fix the type share here to avoid identification challenges that are unrelated to the argument in this paper. When simulating data from the model, we assigned types to each individual according to $\tau_i = 1 + 1(u_i > 1/2)$, where u_i is drawn from the standard uniform distribution.

Figure 6: Approximation performance: model with heterogeneity

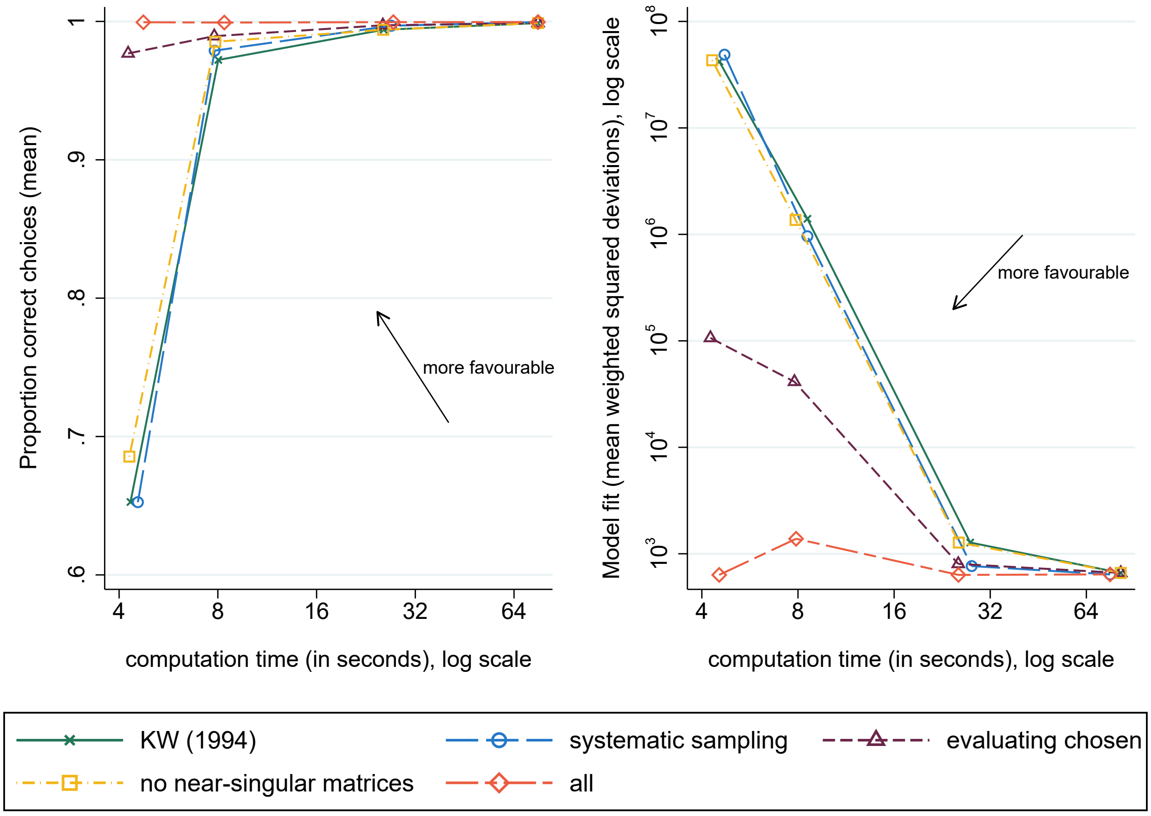


Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density was calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ and types are held constant only for the evaluation of the proportion of correct choices (left panel).

We evaluate the performance of the different varieties of the KW94 method using the same metrics as above. Figure 6 shows the results when at most $M = 500$ nodes are evaluated by Monte Carlo integration. On the whole, the KW94 method works exceptionally well in the model with unobserved heterogeneity. Even using the basic method, around 99 percent of choices are the same in the approximately solved model in most cases. However, in a minority of cases, the basic KW94 approximation — as well as the variations with systematic sampling and data-guided node selection — can go badly awry, resulting in a low proportion of correct choices and very poor fit.

Virtually all of these 'approximation failures' appear to be due to multicollinearity in the interpolation regression. Dispensing with interpolation when the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular removes almost all of them. The version of the KW94 method incorporating all three improvements achieves excellent fit and a near perfect proportion of correctly predicted choices. The explanation for this extremely good performance — even compared to the same method in the model without heterogeneity — is that in the model with heterogeneity, fewer individuals are on the margin between different options, so the heterogeneity to a large extent 'pre-determines' agents' paths through the model.

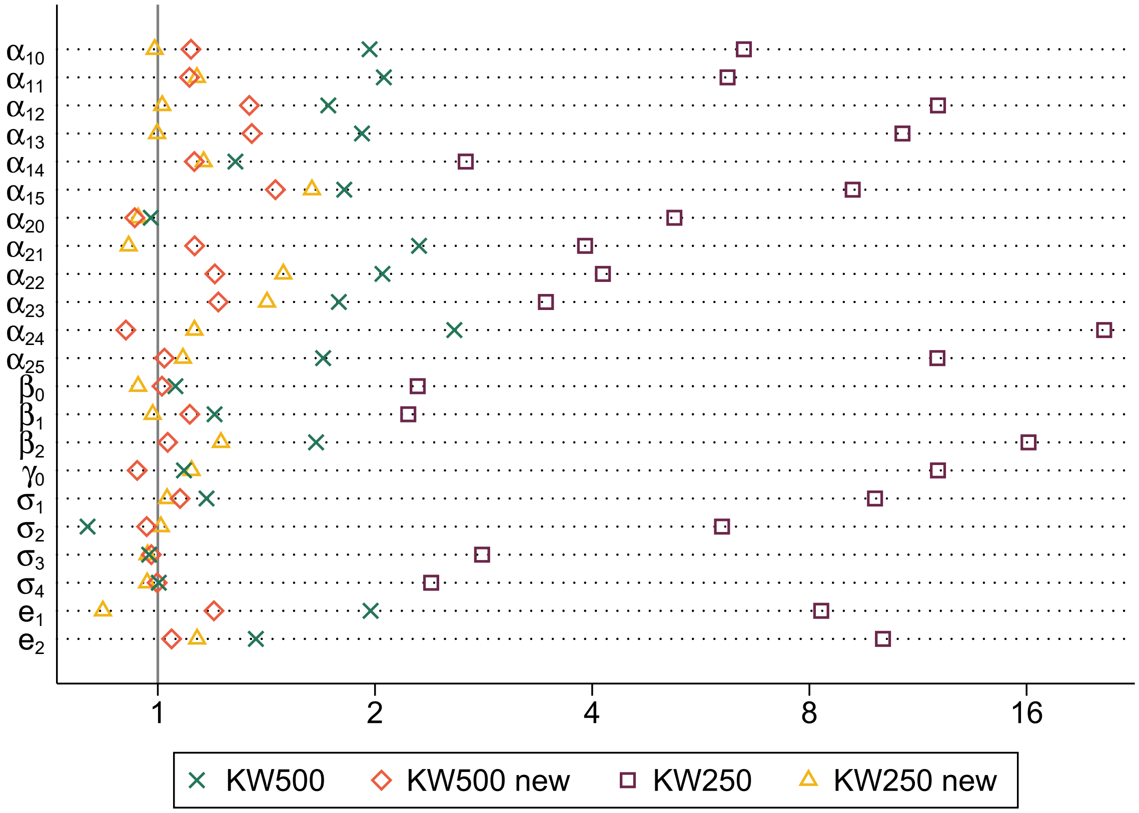
Figure 7: Trade-off between accuracy and computation time: model with heterogeneity



Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Figure 7 shows how the mean fit in the model with heterogeneity varies as the number of points evaluated by Monte Carlo integration is increased. Notably, the full solution provides essentially perfect fit, whether or not systematic sampling is used for numerical integration. Evaluating the chosen nodes still substantially flattens the trade-off between accuracy and computation time. Again the trade-off virtually disappears (for the moderate numbers of evaluation points shown) when all three improvements are combined.[29]

Figure 8: Root mean squared error of parameter estimates relative to full solution: model with heterogeneity



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 15 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated datasets; the 15 estimated parameter sets (50%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated dataset used for estimation was a balanced panel of $N = 2,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 30 steps of the BHHH algorithm were performed. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

5.2 Estimation

One preliminary challenge in estimating this model is that it is only set-identified (in a set of two elements). The reason is that for any given parameterization, a parameterization in which type 2 had the same payoffs as type 1 in the given parameterization and *vice versa* is empirically equivalent. In particular, it is easy to verify that any two parameterizations *a* and

[29] The slightly worse fit in terms of weighted squared deviations when at most $M = 500$ state space points are evaluated is due to the influence of a single outlier.

b are empirically equivalent if $\zeta_1^b = -\zeta_1^a$, $\zeta_2^b = -\zeta_2^a$, $\alpha_{10}^b = \alpha_{10}^a + \zeta_1^a$, $\alpha_{20}^b = \alpha_{20}^a + \zeta_2^a$, and all other parameters are identical. We address this challenge by reparameterizing the solution to the equivalent parameterization if and only if $\hat{\zeta}_1 < 0$ and $\hat{\zeta}_2 < 0$. The log likelihood function now takes the form

$$l(\theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{J} P(\tau_i = j) f(\boldsymbol{d}_{i1}, w_{i1} | \tau_i = j) \prod_{t=2}^{T} f(\boldsymbol{d}_{it}, w_{it} | (\boldsymbol{d}_{i,s})_{s=1}^{s=t-1}, \tau_i = j). \tag{18}$$

Table 2: \$500 tuition fee subsidy: model with heterogeneity

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-.141	-.009	-.138	-.374	-.141	-.147	-.137
	(.005)	(.001)	(.007)	(.517)	(.008)	(.013)	(.007)
White Collar	-.261	-.221	-.268	-.049	-.265	-.279	-.268
	(.006)	(.005)	(.007)	(.473)	(.006)	(.010)	(.010)
School	.403	.248	.406	.450	.406	.427	.405
	(.008)	(.005)	(.010)	(.107)	(.010)	(.013)	(.011)
Home	0	-.018	0	-.027	0	0	0
	(0)	(.001)	(0)	(.045)	(0)	(0)	(0)

Notes. Estimated impact of a \$500 tuition subsidy on average years spent in each occupation. Two balanced panels are simulated from the model for each of 15 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each dataset contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration are different from those used in estimation. Simulations are performed using the full solution method in all cases.

Figure 8 provides a measure of the performance of the different approximation methods when the model with heterogeneity is estimated using Simulated Maximum Likelihood. Estimation was performed exactly as for the model without heterogeneity, except that 30 (instead of 15) steps of the BHHH algorithm were performed and half (instead of 20%) of estimates were dropped to guard against outliers and convergence problems, reflecting the more complex numerical optimization problem.[30] As in the model without heterogenity, our three improvements lead to more accurate parameter estimates that are roughly comparable to those obtained using the full solution. However, the results are noisier, which is partly explained by the lower number of estimation runs used in calculating the root mean squared error.

Table 2 again shows the impact of a \$500 tuition subsidy. The relative performance of the traditional and the new versions of the KW94 method are similar to the canonical model. Notably, the true effect of the tuition subsidy is much smaller in the model with *ex ante* hetero-

[30] The starting vales for the additional model parameters ζ_1 and ζ_2 reflecting the heterogeneous payoffs for different types were set to zero.

geneity, even though the choice shares are similar. Again, the reason is that fewer individuals are on the margin between different options.

6 Conclusion

In this paper, we have suggested three improvements to the KW94 method, which until recently was commonly used by applied researchers to estimate finite-horizon discrete choice dynamic programming problems. First, systematic sampling substantially improves approximation performance. Second, drawing state space points for evaluation from the subset that agents reach in the data is much better than choosing them at random. Third, especially in models with *ex ante* heterogeneity, it is advisable to check for multicollinearity in the interpolation regression and dispense with polynomial interpolation when multicollinearity is detected.

With these improvements, the KW94 method achieves excellent approximation performance even in a model with a large state space and substantial *ex ante* heterogeneity. This suggests that the method has been abandoned prematurely. Its flaws are straightforward to address, and the potential savings in computation time make the effort worthwhile. Holding computation time constant, these improvements translate into more accurate parameter estimates and policy simulations; for a given error tolerance, a larger class of models can be estimated.

Future work is likely to improve on all three of our suggestions. More sophisticated systematic sampling techniques are likely to offer even larger variance reduction in Monte Carlo integration. Other systematic ways of choosing state space points to evaluate by numerical integration may improve on our approach. Other procedures for dealing with multicollinearity may also be superior to our technique; a particularly promising approach may be ridge regression (Hoerl and Kennard, 1970). Most importantly, it will be crucial to test our methods in other contexts besides the canonical KW94 model and its close cousins.

References

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses." *The Quarterly Journal of Economics*. Forthcoming.

Altonji, Joseph G, Anthony A Smith Jr, and Ivan Vidangos. 2013. "Modeling earnings dynamics." *Econometrica*, 81(4): 1395–1454.

Arcidiacono, Peter, and Robert A Miller. 2011. "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity." *Econometrica*, 79(6): 1823–1867.

Belzil, Christian, Jorgen Hansen, and Xingfei Liu. 2017. "Dynamic skill accumulation, education policies, and the return to schooling." *Quantitative Economics*, 8(3): 895–927.

Berndt, Ernst R, Bronwyn H Hall, Robert E Hall, and Jerry A Hausman. 1974. "Estimation and inference in nonlinear structural models." In *Annals of Economic and Social Measurement, Volume 3, number 4*. 653–665. NBER.

Blau, David M, and Donna B Gilleskie. 2006. "Health insurance and retirement of married couples." *Journal of Applied Econometrics*, 21(7): 935–953.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and Jonathan Shaw. 2016. "Female labor supply, human capital, and welfare reform." *Econometrica*, 84(5): 1705–1753.

Blundell, Richard, Monica Costa-Dias, David Goll, and Costas Meghir. 2021. "Wages, Experience, and Training of Women over the Life Cycle." *Journal of Labor Economics*, 39(S1): S275–S315.

Bruins, Marianne, James A Duffy, Michael P Keane, and Anthony A Smith Jr. 2018. "Generalized indirect inference for discrete choice models." *Journal of Econometrics*, 205(1): 177–203.

Crawford, Gregory S, and Matthew Shum. 2005. "Uncertainty and learning in pharmaceutical demand." *Econometrica*, 73(4): 1137–1173.

Eckstein, Zvi, and Kenneth I Wolpin. 1989. "Dynamic labour force participation of married women and endogenous work experience." *The Review of Economic Studies*, 56(3): 375–390.

Eisenhauer, Philipp. 2019. "The approximate solution of finite-horizon discrete-choice dynamic programming models." *Journal of Applied Econometrics*, 34(1): 149–154.

Eisenhauer, Philipp, James J Heckman, and Stefano Mosso. 2015. "Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments." *International Economic Review*, 56(2): 331–357.

Erdem, Tülin, and Michael P Keane. 1996. "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets." *Marketing Science*, 15(1): 1–20.

Geweke, John. 1996. "Monte Carlo simulation and numerical integration." *Handbook of Computational Economics*, 1: 731–800.

Geweke, John, and Michael Keane. 2000. "Bayesian inference for dynamic discrete choice models without the need for dynamic programming." *Simulation-based inference in econometrics: methods and applications*, 100–131.

Hajivassiliou, Vassilis. 2000. "Some practical issues in maximum simulated likelihood." *Simulation-Based inference in econometrics: Methods and Applications*, 71–99.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud. 1996. "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results." *Journal of Econometrics*, 72(1-2): 85–134.

Halton, John H. 1964. "Algorithm 247: Radical-inverse quasi-random point sequence." *Communications of the ACM*, 7(12): 701–702.

Heckman, James J. 1979. "Sample selection bias as a specification error." *Econometrica*, 153–161.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12(1): 55–67.

Hotz, V Joseph, and Robert A Miller. 1993. "Conditional choice probabilities and the estimation of dynamic models." *The Review of Economic Studies*, 60(3): 497–529.

Imai, Susumu, and Michael P Keane. 2004. "Intertemporal labor supply and human capital accumulation." *International Economic Review*, 45(2): 601–641.

Keane, Michael P, and Kenneth I Wolpin. 1994. "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence." *Review of Economics and Statistics*, 648–672.

Keane, Michael P, and Kenneth I Wolpin. 1997. "The career decisions of young men." *Journal of Political Economy*, 105(3): 473–522.

Keane, Michael P, and Kenneth I Wolpin. 2001. "The effect of parental transfers and borrowing constraints on educational attainment." *International Economic Review*, 42(4): 1051–1103.

Keane, Michael P, and Kenneth I Wolpin. 2010. "The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, hispanic, and white women." *International Economic Review*, 51(3): 851–892.

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

Kristensen, Dennis, Lars Nesheim, and Á de Paula. 2015. "CCP and the estimation of nonseparable dynamic models." Mimeo, University College London.

Lee, Donghoon, and Kenneth I Wolpin. 2006. "Intersectoral labor mobility and the growth of the service sector." *Econometrica*, 74(1): 1–46.

Lee, Donghoon, and Kenneth I Wolpin. 2010. "Accounting for wage and employment changes in the US from 1968–2000: A dynamic model of labor market equilibrium." *Journal of Econometrics*, 156(1): 68–85.

Llull, Joan. 2018. "Immigration, wages, and education: A labour market equilibrium structural model." *The Review of Economic Studies*, 85(3): 1852–1896.

McFadden, Daniel. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration." *Econometrica*, 57(5): 995.

Morten, Melanie. 2019. "Temporary migration and endogenous risk sharing in village india." *Journal of Political Economy*, 127(1): 1–46.

Nadarajah, Saralees, and Samuel Kotz. 2008. "Exact distribution of the max/min of two Gaussian random variables." *IEEE Transactions on very large scale integration (VLSI) systems*, 16(2): 210–212.

Sauer, Robert M. 2015. "Does it pay for women to volunteer?" *International Economic Review*, 56(2): 537–564.

Skira, Meghan M. 2015. "Dynamic wage and employment effects of elder parent care." *International Economic Review*, 56(1): 63–93.

Stock, James H, and David A Wise. 1990. "Pensions, the Option Value of Work, and Retirement." *Econometrica*, 1151–1180.

Todd, Petra E, and Kenneth I Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American Economic Review*, 96(5): 1384–1417.

Train, Kenneth E. 2009. *Discrete choice methods with simulation.* Cambridge University Press.

Ulyssea, Gabriel. 2018. "Firms, informality, and development: Theory and evidence from Brazil." *American Economic Review*, 108(8): 2015–47.

Van der Klaauw, Wilbert, and Kenneth I Wolpin. 2008. "Social security and the retirement and savings behavior of low-income households." *Journal of Econometrics*, 145(1-2): 21–42.

Appendix

A Canonical model: Simulated Method of Moments estimation

It is increasingly popular for applied researchers to estimate dynamic discrete choice models using the Simulated Method of Moments.[31] This can be advantageous, for example, when very large datasets are used, multiple datasets are combined, or evaluation of the likelihood is prohibitively complex. As demonstrated in this section, our results apply regardless of whether Simulated Maximum Likelihood or the Simulated Method of Moments is used.

We implement the Simulated Method of Moments approach using our second measure of model fit given in (14) as the criterion function.[32] Unlike in section 3.4, we do not simulate whole paths through the model, but in each period update the parameter vector with observed choices in the data. As a result, individuals in the simulated dataset will in each period have the same vector of pre-determined state variables $\overline{\boldsymbol{S}}_t$ as individuals in the 'true' dataset, as the vector of pre-determined state variables in the canonical KW94 model only depends on observed choices. This approach has the computational advantage that small changes in model parameters typically only affect an individual's decision in a single period rather than in all subsequent periods, as future state variables are unaffected.[33] It is further useful to replace the simulated choices by smooth analogues in order to obtain an entirely smooth objective surface.[34] Then fast and easily parallelized gradient-based optimization routines can be used.[35]

Estimation performance. Figure 9 shows the estimation performance of different versions of the KW94 method relative to the full solution when the Simulated Method of Moments is used in estimation. Again estimation performance is measured by the root mean squared error relative to the full solution. Only the traditional KW94 method with $M = 250$ draws is clearly inferior to the full solution, with all other methods offering comparable performance.

[31] Important recent papers using this method include Altonji, Smith Jr and Vidangos (2013), Blundell et al. (2016), Llull (2018), Ulyssea (2018), Morten (2019) and Abaluck and Adams-Prassl (2021).

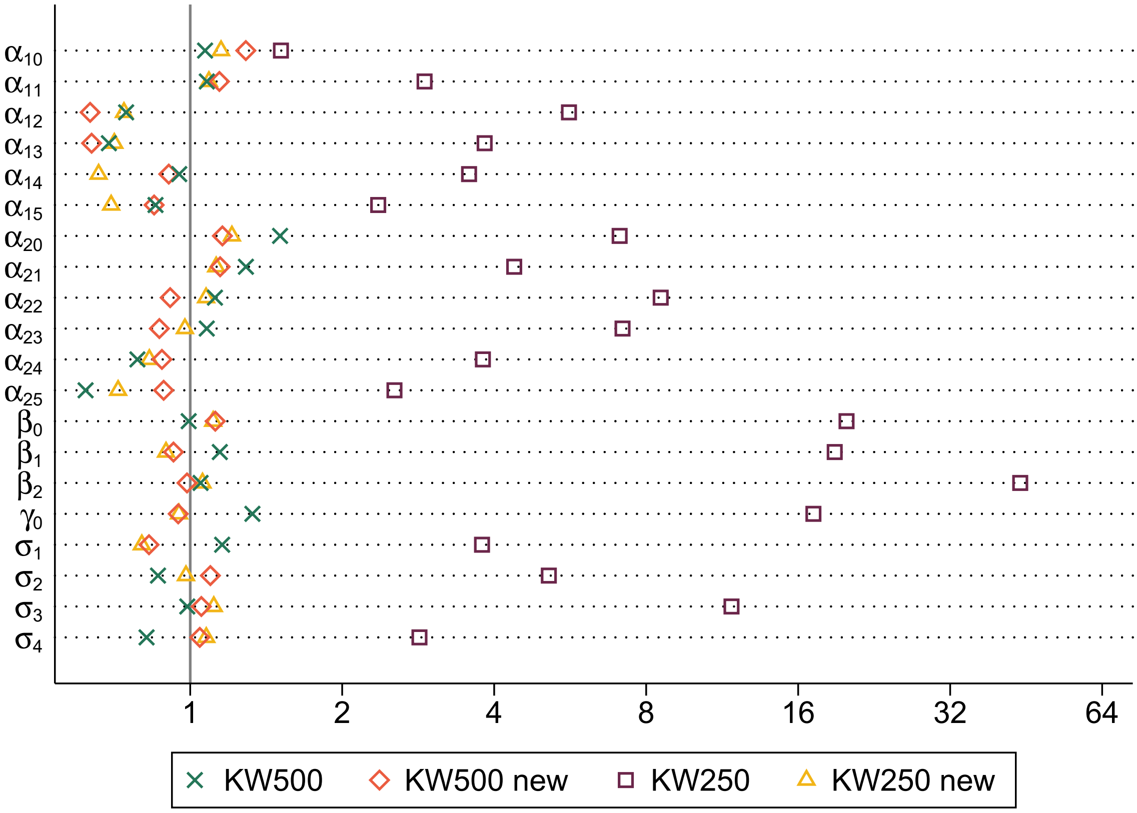
[32] In addition to the moments listed in section 3.4, we also include the entries of transition matrices between choices in each period as moments to aid identification.

[33] Skira (2015) takes a similar approach.

[34] As shown by Bruins et al. (2018), the SMM estimator will still be consistent as the smoothing parameter goes to zero.

[35] We have found the Levenberg-Marquardt algorithm to be the most efficient.

Figure 9: SMM: root mean squared error relative to full solution



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 24 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated datasets; the 6 estimated parameter sets (20%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated dataset used for estimation was a balanced panel of $N = 10,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 15 steps of the Levenberg-Marquardt algorithm were performed.

Table 3: $500 Tuition Fee Subsidy: Simulated Method of Moments

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-3.34	-.126	-3.604	-2.624	-3.634	-3.58	-3.662
	(.119)	(.010)	(.178)	(1.315)	(.206)	(.314)	(.164)
White Collar	2.079	-.235	2.243	1.79	2.257	2.355	2.298
	(.109)	(.013)	(.173)	(.884)	(.181)	(.366)	(.166)
School	1.461	.408	1.552	.984	1.57	1.424	1.56
	(.026)	(.005)	(.054)	(.536)	(.061)	(.063)	(.034)
Home	-.199	-.048	-.191	-.150	-.193	-.199	-.195
	(.011)	(.003)	(.015)	(.087)	(.015)	(.027)	(.018)

Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Two balanced panels are simulated from the model for each of 24 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each dataset contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration are different from those used in estimation. In all cases, simulations were performed using the full solution method.

Policy experiment. Table 3 shows how these estimation results translate into performance in a simple policy experiment. As when parameters are estimated by Simulated Maximum Likelihood, simulation based on parameters estimated using the full solution comes close to simulation results obtained using the true parameters, although there is again a small bias that is very similar in magnitude and direction to the bias obtained when using Simulated Maxmimum Likelihood. The relative performance of the improved and traditional KW94 methods is similar to Simulated Maximum Likelihood.

B Further analytical bounds for the Keane and Wolpin method

While there is in general no analytical solution available for the continuation values in finite-horizon discrete choice dynamic programming models, it is generally possible to establish a lower bound. A basic lower bound that forms the basis of the KW94 approximation method is Jensen's Inequality. In the notation of section 3.4, imposing this bound is the same as requiring that $g(.) \geq 0$.

Further analytical bounds are often available. One avenue for deriving tighter bounds relies on the properties of the max function combined with Jensen's Inequality. For instance,

$$\begin{aligned} \mathrm{E}\left[\max(V_1, V_2, V_3, V_4)\right] &= \mathrm{E}\left[\max\left[\max(V_1, V_2), \max(V_3, V_4)\right]\right] \\ &\geq \max\left[\mathrm{E}\left[\max(V_1, V_2)\right], \mathrm{E}\left[\max(V_3, V_4)\right]\right]. \end{aligned} \tag{B.1}$$

The lower bound in the second line of (B.1) is weakly tighter than the bound provided by Jensen's inequality. This is helpful because it is often possible to derive an analytical form or a tighter bound for the expectation of the maximum of two options. In the canonical model of KW94, an analytical form is available for $\mathrm{E}\left[\max(V_3, V_4)\right]$ and a tighter bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$ (taking future continuation values as given).

The analytical form for $\mathrm{E}\left[\max(V_3, V_4)\right]$ can be derived from the fact that V_3 and V_4 are jointly Normally distributed in the KW94 model. If $V_3 \sim N(\mu_3, \sigma_3^2)$, $V_4 \sim N(\mu_4, \sigma_4^2)$ and the correlation of V_3 and V_4 is ρ_{34}, then the expectation of the maximum of V_3 and V_4 takes the known form (see e.g. Nadarajah and Kotz, 2008):

$$\mathrm{E}\left[\max(V_3, V_4)\right] = \mu_3 \Phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) + \mu_4 \Phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) + \theta \phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) \tag{B.2}$$

where $\theta = \sqrt{\sigma_3^2 + \sigma_4^2 - 2\rho_{34}\sigma_3\sigma_4}$ and Φ and ϕ are, respectively, the cdf and the pdf of the standard Normal distribution.

A tighter bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$ can be derived from the fact that for $k = 1, 2$, V_{kt} can be decomposed into $V_{kt} = u_{kt} + \delta\,\mathrm{E}(V_{t+1}^k)$. Hence

$$\mathrm{E}\left[\max(V_{1t}, V_{2t})\right] \geq \mathrm{E}\left[\max(u_{1t}, u_{2t})\right] + \delta \min\left[\mathrm{E}(V_{t+1}^1), \mathrm{E}(V_{t+1}^2)\right]. \tag{B.3}$$

As there is no continuation value in the final period, the second term drops out and the expression holds with equality if $t = T$.

u_{1t} and u_{2t} are jointly lognormally distributed, and there is an analytical form available for the expectation of the maximum of two correlated lognormal random variables. In particular, it can be shown that if $\log u_1 \sim N(\mu_1, \sigma_1^2)$, $\log u_2 \sim N(\mu_2, \sigma_2^2)$ and the correlation of $\log u_{1t}$ and $\log u_{2t}$ is ρ_{12}, then the expecation of the maximum of u_{1t} and u_{2t} takes the form:

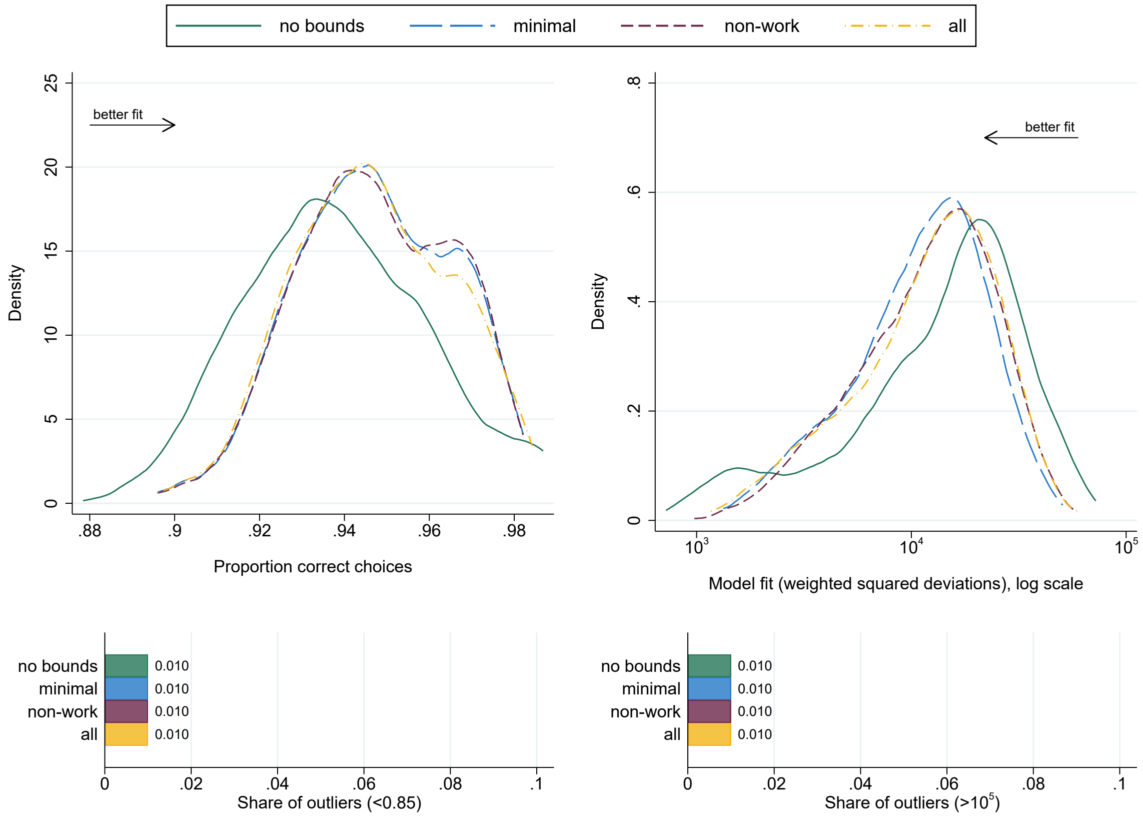
$$\begin{aligned}\mathrm{E}\left[\max(u_{1t}, u_{2t})\right] = {} & \exp\left(\mu_1 + \frac{\sigma_1^2}{2}\right)\Phi\left(\frac{\mu_1 - \mu_2 + \sigma_1^2 - \rho_{12}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 + \sigma_2^2}}\right) \\ & + \exp\left(\mu_2 + \frac{\sigma_2^2}{2}\right)\Phi\left(\frac{\mu_2 - \mu_1 + \sigma_2^2 - \rho_{12}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 + \sigma_2^2}}\right).\end{aligned} \tag{B.4}$$

Figure 10 shows what happens to the simulation accuracy of the traditional KW94 method when different bounds are imposed. The column labelled "no bounds" shows the fit when no bounds are imposed at all. "Minimal" shows the effect of imposing the bound that

$$\underbrace{\mathrm{E}[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)]}_{\text{EMAX}} \geq \underbrace{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\text{MAXE}},$$

which follows from a straightforward application of Jensen's Inequality. "Non-work" additionally imposes the bound based on the analytical expression for $\mathrm{E}\left[\max(V_3, V_4)\right]$. "All" also imposes the additional bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$.

Figure 10: Performance of the KW94 method with different bounds



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "No bounds" shows the fit when no bounds are imposed at all. "Minimal" shows the effect of imposing the bounds that $EMAX \geq MAXE$ that follows from a straightforward application of Jensen's Inequality. "Non-work" additionally imposes the bound based on the analytical expression for $\mathrm{E}\left[\max(V_3, V_4)\right]$. "All" also imposes the additional bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

On the whole, the improvements in fit to be gained from the additional bounds in the canonical model appear to be minor. Only the "minimal" bound that $EMAX \geq MAXE$ leads to a substantial improvement in approximation performance. On the whole, the effect of imposing bounds is not unambiguously positive. One reason is likely to be that what counts for approximation performance is not how closely the approximation reproduces the true EMAX in absolute terms, but only the differences between the nodes. Hence an increase in the approximated value of some nodes but not others as a result of imposing a tighter lower bound can be counterproductive. In sum, it may be best for applied researchers to stick to the "minimal" approach to bounding recommended in KW94.

C Alternative approximation methods using MAXE

Alternatives to the KW94 approximation method exist and have been used in empirical work. They can be broadly divided into two groups: those that do and do not rely on the MAXE approximation. In this appendix, we present evidence on the performance of other methods that, like the KW94 approximation, rely on the MAXE approximation. Alternative approximation methods that do not rely on the MAXE approximation include Geweke and Keane (2000) and Sauer (2015) (also used in Belzil, Hansen and Liu, 2017).[36]

All methods relying on the MAXE approximation are built around the observation that in many models, $\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]$ ('MAXE') provides a decent approximation to the actual expected value of each state space point $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX'). At the same time, MAXE is usually very cheap to calculate, as no numerical integration is required. As a result, these methods can allow for very fast approximation even when the state space is orders of magnitude larger than in the canonical KW94 model.

The most common approach is to simply replace EMAX by MAXE in the computation of continuation values (e.g. Stock and Wise, 1990). This works well in some contexts, but fails in the canonical model: on average, not even a third of choices accord with the 'true' simulated choices. The reason is that differences between EMAX and MAXE are a key driver of decisions in that model, so approximating EMAX by MAXE leads to very different simulated choices.

A potential improvement on this approach resembles the KW94 method. Under this procedure, one approximates the option value of state space points by introducing an auxiliary polynomial into the model representing the difference between MAXE and EMAX. In particular, in the canonical model, we replace each $\hat{\mathrm{E}}\left[V_t(\boldsymbol{S}_t)|\overline{\boldsymbol{S}}_t, d_{kt}=1\right]$ by

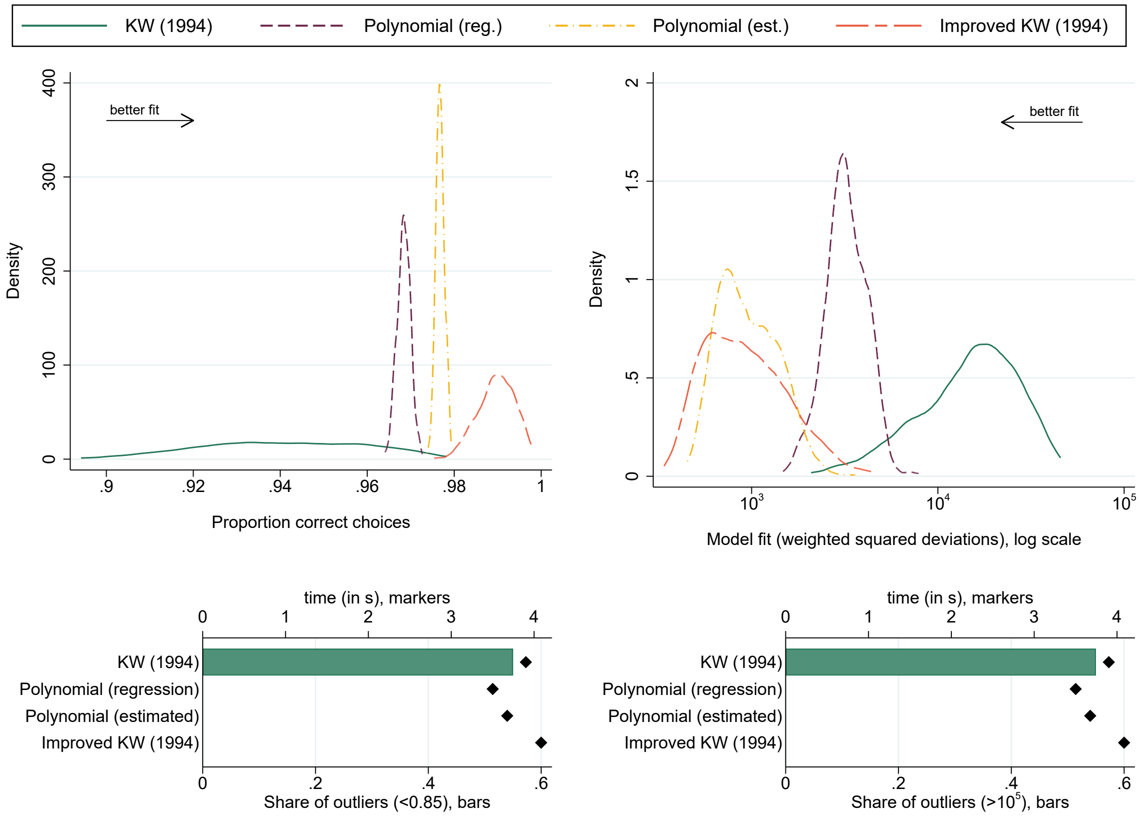
$$\begin{aligned}\check{\mathrm{E}}\left[V_t(\boldsymbol{S}_t)|\overline{\boldsymbol{S}}_t, d_{kt}=1\right] &= \max_{k\leq K} \mathrm{E}\left[V_{kt}(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t)\right] \\ + 1(s_t < s_{max}) &\left(\sum_{j=1}^{K} \pi_j \left[\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right] - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right]\right] + \sum_{j=1}^{K} \pi_{K+j} \sqrt{\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right] - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right]}\right) \\ &+ \pi_{2K+1} 1(s_t = s_{max}). \quad \text{(C.1)}\end{aligned}$$

The $2K+1$ elements of $\boldsymbol{\pi}$ are then estimated along with the vector of structural parameters

[36] A different framework entirely is the Conditional Choice Probability (CCP) approach (Hotz and Miller, 1993; Arcidiacono and Miller, 2011). This approach generally requires errors in the model to be additive (although see Kristensen, Nesheim and de Paula, 2015, for arguments why this condition can in principle be relaxed). This is not the case in either of the models discussed in this paper, as the wage functions take a semi-log form.

$\boldsymbol{\theta}$.[37]

Figure 11: Alternative methods using MAXE: canonical model



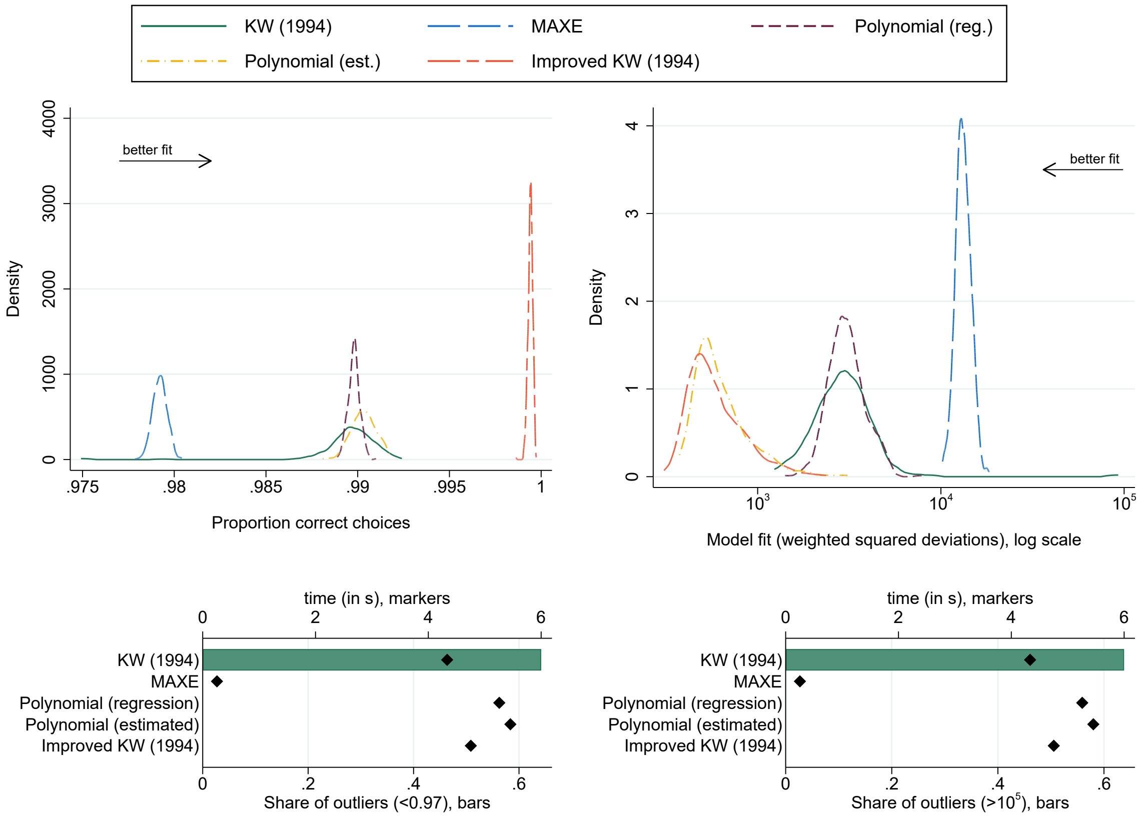
Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Polynomial (reg.)" shows the performance of the alternative approach if the auxiliary parameter vector π is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') as the dependent variable and including one observation for each person and period in the simulated data. "Polynomial (est.)" shows the fit when $\boldsymbol{\pi}$ is estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Figure 11 shows the approximation performance of this method compared to both the traditional version and our improved version of the KW94 method (with a maximum of $M = 250$ nodes evaluated to roughly match the other method in terms of computation time). "Polynomial (regression)" shows the performance of the alternative approach if the auxiliary parameter vector $\boldsymbol{\pi}$ is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$

[37] Note that we do not include an intercept term or functions of the time period t. As the estimation of $\boldsymbol{\pi}$ is solely based on observed choices, these 'scale' parameters are — depending on how they are included — either not identified at all or only very weakly identified as a result of the maximum schooling limit.

('EMAX') as the dependent variable and including one observation for each person and period in the simulated data.[38] "Polynomial (estimated)" shows the fit when $\boldsymbol{\pi}$ is estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values.

Figure 12: Alternative methods using MAXE: model with heterogeneity



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Polynomial (reg.)" shows the performance of the alternative approach if the auxiliary parameter vector $\boldsymbol{\pi}$ is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') as the dependent variable and including one observation for each person and period in the simulated data. "Polynomial (est.)" shows the fit when $\boldsymbol{\pi}$ is estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

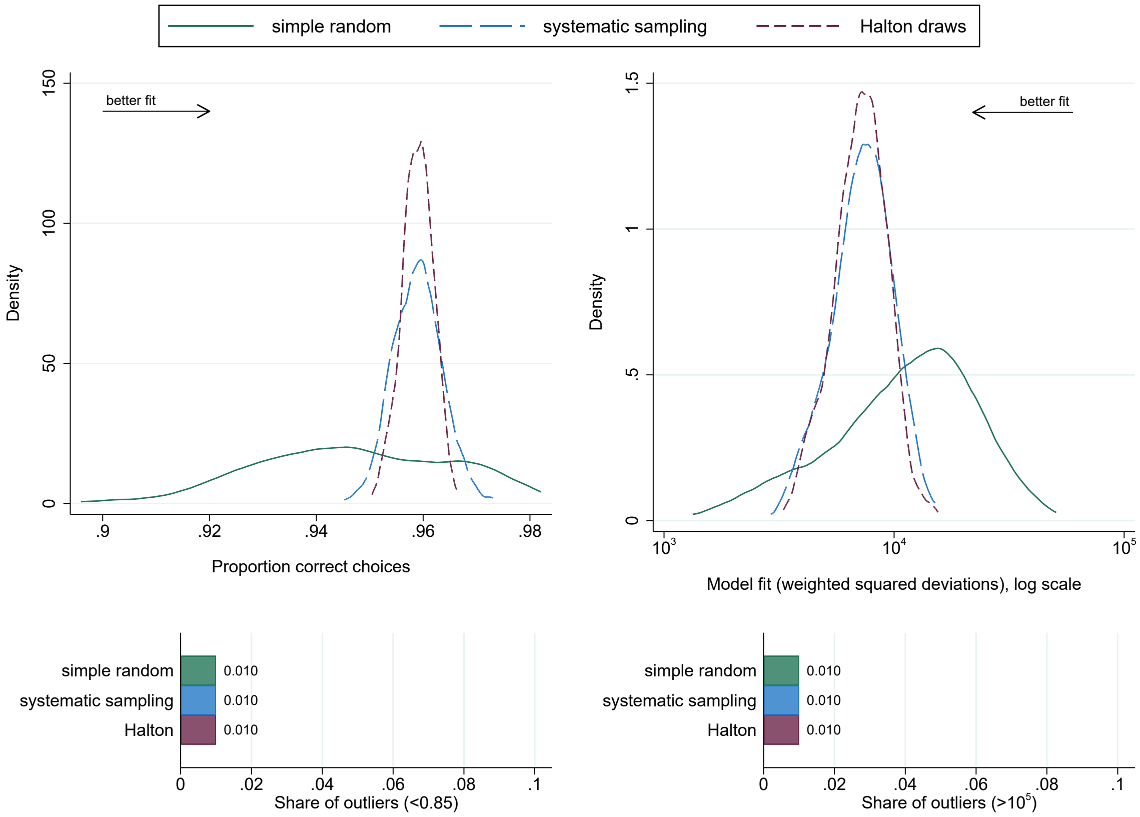
Overall, the alternative approach is no match for the improved KW94 method, even when the parameters of the auxiliary parameters are estimated from the 'true' EMAX values, or holding constant the structural parameters at their true values (either of which would be impossible

[38] As scale is not identified (see fn. 37 above), we also include a full set of period dummies in that regression, but disregard the estimated period effects.

in practice). However, under these idealized conditions, the alternative approach does offer better approximation performance at similar computational cost than the traditional KW94 method with at most $M = 250$ evaluation points. In part, this is because the KW94 method performs poorly in over half of all cases when only $M = 250$ nodes are evaluated using Monte Carlo integration.

Figure 12 is the equivalent figure for the model with heterogeneity. We include the approximation of EMAX by MAXE in this figure, as in the model with heterogeneity, that approximation also offers acceptable performance. All methods perform much better, as more *ex ante* heterogeneity means that the precise continutation value of each state space point matters less in shaping agents' decisions.

Figure 13: Performance of the KW94 method with different sampling methods



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Simple" refers to simple random sampling, "Systematic" refers to the systemantic sampling algorithm outlined in section 3.1, and Halton draws are obtained following the algorithm from Halton (1964). The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

The improved KW94 method still dominates in terms of approximation performance with a virtually perfect fit. The method with an auxiliary polynomial again yields better approximation performance than the traditional KW94 method. The simple MAXE method leads to a somewhat worse fit, but is much faster to compute. This suggests that in models with extremely large state spaces and substantial heterogeneity, it might be advisable to estimate the model using the simple MAXE approximation in an initial step.[39]

D Comparison of different sampling methods

D.1 KW94 method

Figure 13 shows the influence of different sampling methods on the performance of the traditional version of the KW94 method in the canonical model with a maximum of $M = 500$ evaluation points. Simple random draws are compared to the systematic sampling method outlined in section 3.1 and Halton (1964) systematic draws. Both systematic sampling and Halton draws lead to large improvements in approximation accuracy. Halton draws are slightly better than systematic sampling, but the differences are minor.[40]

D.2 Full solution

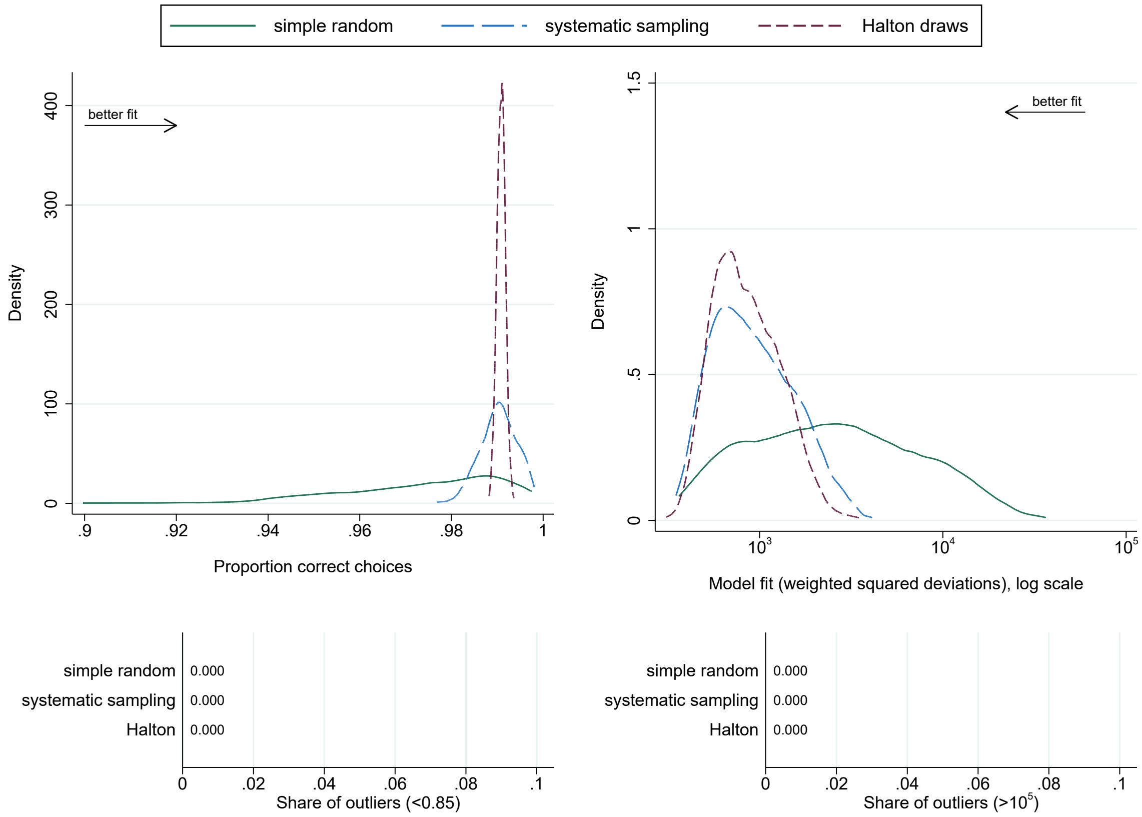
Figure 14 shows the influence of different sampling methods on the performance of the full solution method in the canonical KW94 model. Accuracy gains are even larger than when using the KW94 method, with both systematic sampling and Halton draws offering an excellent fit. Using the full solution method, the advantage of Halton draws is somewhat larger in relative terms.[41]

[39] Whether the method with an auxiliary polynomial has a useful role to play in some contexts is an interesting question for futher research.

[40] Intuitively, the reason that Halton draws perform somewhat better is that they ensure even integral coverage in multiple dimensions, whereas our systematic sampling method only ensures even coverage in each dimension individually.

[41] Intuitively, even integral coverage becomes more important when other sources of approximation error have been eliminated.

Figure 14: Performance of the full solution method with different sampling methods



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Simple" refers to simple random sampling, "Systematic" refers to the systemantic sampling algorithm outlined in section 3.1, and Halton draws are obtained following the algorithm from Halton (1964). The two panels show our two different metrics of approximation performance. Realized shocks $(\epsilon_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

E Bias and standard deviation of estimated parameters

E.1 Canonical model: Simulated Maximum Likelihood

Table 4 shows the average deviation of the estimated parameters from the true parameters. For the full solution, these average deviations are mostly quite small relative to the scale of each parameter. The exception are the standard deviations of the non-work options a_{33} and a_{44}, which are substantially underestimated. This is a known artefact of Simulated Maximum Likelihood estimation in this model; the same bias is found in KW94.[42] Biases for the improved

[42] This is likely related to the use of the smoothed-logit simulator (McFadden, 1989), which is known to introduce bias (cf. the discussion in Keane and Wolpin, 1994). In particular, the smoothed-logit simulator tends to overstate the likelihood of tail events, which may explain a downward bias in the error variance.

versions of the KW94 method are generally very close to the full solution. They are mostly larger for the traditional KW94 method, especially when only a maximum of $M = 250$ state space points are evaluated by Monte Carlo integration.

Table 4: Average deviation (bias) of estimated parameters in the canonical model

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.00464	-.00273	.00409	.00512	.00494
α_{01}	.038	-.00012	.00071	-9.5e-05	-9.4e-05	-.00014
α_{02}	.033	-.00034	-.00057	-.0003	-.00039	-.00035
α_{03}	.0005	-9.1e-06	-1.7e-05	-8.4e-06	-1.0e-05	-9.2e-06
α_{04}	0	-.00019	-.0001	-.00017	-.00049	-.00022
α_{05}	0	-5.1e-06	5.0e-05	-3.4e-06	-1.4e-05	-6.1e-06
α_{10}	8.48	.00392	.01703	.0042	.0056	.0042
α_{11}	.07	-.00016	-.00119	-.00021	-.00035	-.0002
α_{12}	.067	-7.3e-05	.00139	4.2e-05	.00011	1.5e-05
α_{13}	.001	8.4e-08	3.7e-05	3.8e-06	4.8e-06	3.0e-06
α_{14}	.022	-.00018	-.00338	-.00025	-.00045	-.00022
α_{15}	.0005	2.1e-06	-.00013	-1.1e-06	-1.5e-05	5.7e-08
β_0	0	77.502	-39.536	84.26	-3.8551	90.82
β_1	0	31.616	101.55	34.448	27.902	37.208
β_2	4000	-131.72	-1500.5	-137.01	-159.78	-131.77
γ_{0_f}	17750	68.271	54.957	67.283	54.334	67.768
a_{11}	.2	-.00107	-.00098	-.00091	-.00099	-.00104
a_{22}	.25	1.2e-05	-.0005	2.8e-05	-9.0e-05	2.7e-05
a_{33}	1500	-372.81	574.77	-371.67	-378.99	-372.13
a_{44}	1500	-246.99	613.65	-252.8	-249.09	-237.67

Notes. The average deviations from the true parameters were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

Table 5 show the sample standard deviation of the estimated parameters. Sample standard deviations tend to be of similar orders of magnitude as the average deviation from the true parameters. Standard deviation methods for all approximation methods are similar to the full solution, except for the traditional KW94 method with $M = 250$, which has much larger standard deviations. This appears to be driven by a minority of estimation runs that do not converge to parameter values near the true parameters (not shown).

Table 5: Sample standard deviation of estimated parameters in the canonical model

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.0081	.01564	.00837	.00869	.00846
α_{01}	.038	.00071	.0016	.00075	.00075	.00073
α_{02}	.033	.00032	.00069	.0003	.00033	.00032
α_{03}	.0005	1.0e-05	2.0e-05	9.8e-06	1.1e-05	1.0e-05
α_{04}	0	.00034	.00112	.00033	.00028	.00034
α_{05}	0	1.7e-05	8.6e-05	1.7e-05	1.6e-05	1.7e-05
α_{10}	8.48	.0041	.01904	.00386	.00428	.00416
α_{11}	.07	.00028	.00143	.00029	.00028	.00027
α_{12}	.067	.00035	.00172	.00037	.00037	.00035
α_{13}	.001	1.1e-05	4.4e-05	1.2e-05	1.2e-05	1.1e-05
α_{14}	.022	.00014	.0038	.00016	.00014	.00014
α_{15}	.0005	1.2e-05	.00015	1.3e-05	1.2e-05	1.2e-05
β_0	0	132.63	221.73	133.55	142.38	128.2
β_1	0	77.847	279.86	80.672	67.949	78.504
β_2	4000	57.199	1713.6	54.389	58.261	55.63
γ_{0_f}	17750	67.211	68.543	64.427	69.477	68.538
a_{11}	.2	.00121	.00126	.00118	.00123	.00121
a_{22}	.25	.00075	.00163	.00076	.00076	.00076
a_{33}	1500	56.649	1369.9	54.182	55.83	55.731
a_{44}	1500	94.477	1209.9	96.81	100.84	94.769

Notes. The sample standard deviations were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

E.2 Canonical model: Simulated Method of Moments

Tables 6 and 7 are the same tables for the Simulated method of Moments (SMM) estimation. Both average deviations from the truth and sample standard deviations are somewhat larger for SMM, likely reflecting lower efficiency compared to Simulated Maximum Likelihood. The relative performance of different approximation methods is similar to the maximum likelihood results, with the improved KW94 method performing both better and similar to the full solution.

Table 6: Average deviation (bias) of estimated parameters in the canonical model

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	-.00401	-.00487	-.00428	-.00337	-.00404
α_{01}	.038	.00041	.00246	.00037	.0005	.0004
α_{02}	.033	-8.1e-05	-.00136	-2.2e-06	-1.6e-05	-7.8e-05
α_{03}	.0005	-2.1e-06	-2.8e-05	-3.1e-07	-8.1e-07	-3.2e-06
α_{04}	0	-2.4e-05	-.00204	-.00025	-.00016	.00013
α_{05}	0	1.0e-06	-6.8e-05	-1.2e-05	3.0e-06	7.8e-06
α_{10}	8.48	-.00284	.02545	-.00204	.00478	-.00326
α_{11}	.07	.00044	-.00053	.00034	-.00029	.00048
α_{12}	.067	-.00027	-.00119	-.00018	.00011	-.00023
α_{13}	.001	-5.8e-06	-2.3e-05	-3.5e-06	4.1e-06	-5.2e-06
α_{14}	.022	9.7e-05	-.00131	-7.6e-06	-.0002	.0001
α_{15}	.0005	3.2e-05	-4.7e-05	1.8e-05	8.2e-06	3.5e-05
β_0	0	-69.664	-2378.2	-71.489	10.043	-77.732
β_1	0	135.36	-1598.7	81.936	50.341	144.48
β_2	4000	-136.09	3561	-156.03	-156.2	-150.1
γ_{0_f}	17750	129.96	-1923.7	171.41	-113.5	132.29
a_{11}	.2	.00024	-.00623	.00065	-.00149	3.2e-05
a_{22}	.25	-.00133	-.00335	-.0011	-.00055	-.00157
a_{33}	1500	-493.64	2968.3	-548.88	-490.34	-528
a_{44}	1500	-645.01	1019.1	-751.39	-143.13	-666.45

Notes. The average deviations from the true paramters were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 10,000$ individuals each. $D = 2000$ solution draws were used in model solution.

Table 7: Sample standard deviation of estimated parameters in the canonical model

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.01359	.02088	.01574	.01479	.01782
α_{01}	.038	.00128	.00301	.00142	.00136	.00148
α_{02}	.033	.00042	.00198	.00032	.00032	.00026
α_{03}	.0005	1.4e-05	4.4e-05	9.8e-06	9.5e-06	8.2e-06
α_{04}	0	.00118	.00366	.00074	.00111	.00106
α_{05}	0	6.5e-05	.00014	4.4e-05	5.5e-05	5.4e-05
α_{10}	8.48	.00543	.03507	.00715	.00789	.00631
α_{11}	.07	.00052	.00297	.0007	.00084	.00062
α_{12}	.067	.00033	.00347	.00042	.00047	.00032
α_{13}	.001	1.1e-05	8.6e-05	1.2e-05	1.3e-05	9.4e-06
α_{14}	.022	.00059	.00184	.0005	.00042	.00051
α_{15}	.0005	7.6e-05	.0002	5.7e-05	5.1e-05	6.4e-05
β_0	0	240.19	4366.9	268.42	248.55	269.44
β_1	0	193	4180.2	195.47	266.82	163.09
β_2	4000	101.68	6656.6	88.229	84.553	72.841
γ_{0_f}	17750	235.99	4207.4	188.49	340.39	217.86
a_{11}	.2	.00272	.00816	.00209	.00277	.00227
a_{22}	.25	.00126	.00887	.00143	.0015	.00125
a_{33}	1500	136.79	5370.6	155.29	121.98	105.89
a_{44}	1500	585.27	2280.9	556.62	707.27	619.75

Notes. The sample standard deviations were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 10,000$ individuals each. $D = 2000$ solution draws were used in model solution.

E.3 Model with heterogeneity

Tables 8 and 9 are the analogous tables for the model with heteroegeneity. Both average deviations from the truth and sample standard deviations are comparable to the canonical model. The heterogeneity parameters ζ_1 and ζ_2 are precisely estimated. The relative performance of different approximation methods is similar to the results for the canonical model, with the improved KW94 method performing both better and similar to the full solution.

Table 8: Average deviation (bias) of estimated parameters in the model with heterogeneity

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.00268	-.04773	.00205	.00648	-.00083
α_{01}	.038	-.00013	.00339	-6.3e-05	-.00027	-4.1e-05
α_{02}	.033	-2.9e-05	.00156	-9.6e-05	-8.0e-06	-9.9e-05
α_{03}	.0005	-1.0e-06	2.7e-05	-2.7e-06	-1.2e-06	-2.7e-06
α_{04}	0	-.00051	.00064	-.00055	-.00057	-.00056
α_{05}	0	-7.3e-06	.00013	-1.4e-05	2.7e-06	-1.6e-05
α_{10}	8.48	-.00294	.01362	-.00363	.00556	-.00249
α_{11}	.07	-.00016	-.00044	-8.6e-05	-.00079	-.00026
α_{12}	.067	.00033	.0013	.00047	.00065	.00038
α_{13}	.001	1.0e-05	3.6e-05	1.5e-05	1.9e-05	1.3e-05
α_{14}	.022	-.00019	-.00373	-.00024	-.00065	-.00013
α_{15}	.0005	-6.3e-06	-9.7e-05	-8.7e-06	-2.0e-05	-7.4e-06
β_0	0	20.062	-77.889	-82.077	17.9	-85.364
β_1	0	-32.333	-154.04	-87.964	-89.071	-175.31
β_2	4000	-42.963	-674.81	-60.823	-90.592	-45.135
γ_{0_f}	17750	53.26	517	54.197	54.645	49.307
a_{11}	.2	8.8e-05	.00239	.00024	4.7e-05	.00017
a_{22}	.25	.00084	8.2e-05	.00111	.00029	.00073
a_{33}	1500	-363.2	299.92	-355.42	-354.28	-354.49
a_{44}	1500	-431.49	-44.401	-405.29	-431.6	-430.73
ζ_1	.05	-.00158	-.00449	-.00115	-.00467	.00157
ζ_2	.1	.00325	.00216	.00158	.00268	.00422

Notes. The average deviations from the true parameters were calculated over 15 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

Table 9: Sample standard deviation of estimated parameters in the model with heterogeneity

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.01781	.10606	.01772	.03479	.02002
α_{01}	.038	.00112	.00598	.00127	.0023	.00124
α_{02}	.033	.0002	.00186	.00018	.00035	.00025
α_{03}	.0005	4.5e-06	4.1e-05	3.6e-06	8.7e-06	5.6e-06
α_{04}	0	.00042	.00168	.00054	.00064	.0005
α_{05}	0	2.3e-05	.00018	3.7e-05	4.4e-05	3.1e-05
α_{10}	8.48	.00858	.04519	.00768	.0068	.00806
α_{11}	.07	.00043	.00174	.00041	.00067	.00044
α_{12}	.067	.00029	.00127	.00046	.00061	.00036
α_{13}	.001	1.0e-05	3.6e-05	1.5e-05	1.8e-05	1.3e-05
α_{14}	.022	.00023	.00482	.00023	.00039	.00023
α_{15}	.0005	1.6e-05	.00018	1.6e-05	2.0e-05	1.5e-05
β_0	0	338.93	773.99	307.36	358.7	332.58
β_1	0	330.2	720.72	313.57	386.94	319.6
β_2	4000	55.69	910.25	60.284	71.784	56.798
γ_{0_f}	17750	43.798	659.24	54.776	51.562	41.728
a_{11}	.2	.00082	.00777	.00081	.00096	.00087
a_{22}	.25	.00113	.00864	.00087	.0011	.00115
a_{33}	1500	77.369	1034.9	51.414	70.033	81.722
a_{44}	1500	62.034	1077.8	118.4	70.705	60.681
ζ_1	.05	.00808	.06834	.00681	.01552	.00973
ζ_2	.1	.00601	.06966	.00764	.009	.00572

Notes. The sample standard deviations were calculated over 15 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.