Robustness Checks in Structural Analysis*

Sylvain Catherine† Mehran Ebrahimian‡
Mohammad Fereydounian§ David Sraer¶
David Thesmar‖

January 1, 2026

Abstract

This paper introduces a computationally efficient methodology for estimating variants of structural models. Our approach approximates the relationship between moments and parameters, offering a low-cost alternative to traditional estimation methods. We establish general convergence conditions, and show they are widely applicable, even for models with discontinuities. We also provide convergence rate bounds for Kernel and Neural Net approximations, with the latter demonstrating superior performance in higher dimensions.

We apply this methodology to two standard structural models in financial economics: (1) dynamic corporate finance and (2) life-cycle portfolio choice. We demonstrate the reliability of our approach through simulations and then use it to explore identification, robustness to sample splits, moment selection, and model misspecification. These explorations are computationally infeasible with standard techniques, but become trivial with our methodology.

*We thank participants at seminars and conferences for their comments. We are especially grateful to Hui Chen, Dan Green, Maryam Farboodi, Jonathan Parker, Demian Pouzo and Toni Whited.

†University of Pennsylvania
‡Stockholm School of Economics
§University of Pennsylvania
¶UC Berkeley, NBER and CEPR
‖MIT, NBER and CEPR

1 Introduction

Robustness checks are a standard feature of applied empirical work in economics and finance. After establishing their main results, researchers typically examine whether alternative mechanisms can explain their findings and provide additional analyses to control for such channels. For instance, they reestimate their main specification across various subsamples, either expecting results to be stable or to vary in a specific way. In other cases, they estimate different specifications to account for alternative channels, anticipating that their main predictions will hold. These analyses (sample splits, changing the model) are part of the standard toolbox used by applied researchers.

Structural estimation is an another popular approach in all subfields of financial economics, but robustness checks are rare in this type of research, primarily due to computational constraints. As a first example, consider the case of sub-sample splits, which involve reestimating a model for particular periods or groups of observations. In reduced-form work, this analysis has negligible computational cost. However, in structural research, it can be prohibitively expensive, as each new estimation may take days and often requires human monitoring to fine-tune optimization. Second, consider robustness checks in reduced-form work that introduce additional control variables in regressions. Again, these checks are nearly costless. In simulation-based estimation, a related exercise might involve "purging" moments from control variables and estimating variants of the model fitted against such refined moments. Alternatively, the econometrician may want to evaluate parameter robustness to moment selection, varying the number and nature of moments against which the model is estimated. In many cases, the computational burden of reestimating the structural model implies that very few alternatives, if any, can be considered in practice.

This paper overcomes this challenge by introducing a computationally efficient methodology for estimating variants of structural models. We provide convergence conditions and characterization, and present two applications using popular models in financial economics: life-cycle portfolio choice and corporate finance.

Our approach works as follows. We consider the estimation of a structural model $\mathcal{S}$. Given deep parameters θ, the model generates moments $f(\theta)$. In most applications, such as dynamic choice models, f is calculated numerically. An estimate θ^* of structural parameters is then obtained by minimizing a distance between simulated

moments $f(\theta)$ and empirical moments m:

$$\theta^* = \arg\min_{\theta} (m - f(\theta))^\top W (m - f(\theta)),$$

where W is a weighting matrix. While numerically computing $f(\theta)$ for any given θ can be done reasonably fast with a simulation, the *estimation process* can be computationally costly as it requires a large number of such simulations, as well as human monitoring. This computational burden limits the number of estimations that are feasible for a given research project. This limits the amount of robustness analysis that can be conducted.

We propose to reduce this computational cost by building an *approximation* of f. Importantly, we do not approximate the solution of the *model* itself, as explored in various works.[1] Instead, we directly approximate the *moment function* f.

Our methodology follows three main steps. First, we generate a *training dataset* $\mathcal{D}_n$, comprising n (vectors of) parameters values and their corresponding moment values. This training dataset is fixed once and for all. This step is computationally intensive since n can be large. However, it is *not* more costly than a single model estimation, which also requires a large number of simulations. In the second step, we fit an approximate moment function f_n. Consistent with findings in the literature, we find that Neural Nets perform particularly well, especially when θ has a large dimension. This second step is computationally inexpensive. Finally, in the third step, we estimate parameters by matching moments using the approximate function f_n instead of the true function f:

$$\hat{\theta}_n = \arg\min_{\theta} (m - f_n(\theta))^\top W (m - f_n(\theta)).$$

This step incurs almost no computational cost since f_n is already known.

In Section 3, we examine the general applicability of our method. First, we provide conditions under which the approximate estimate $\hat{\theta}_n \to \theta^*$ as the training data size $n \to \infty$. A key condition is the continuity of moments f with respect to θ (which does not imply that the *model* itself is continous – more on this below). Under slightly

[1] See for example Fernandez-Villaverde et al. (2021), Fernández-Villaverde et al. (2021), and Duarte (2020).

stronger regularity conditions, we derive a formula for the *approximation error*:

$$\hat{\theta}_n - \theta^* \approx \underbrace{\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)\right)^{-1} \nabla f_n(\hat{\theta}_n)^\top W}_{\equiv \Lambda_n} \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right),$$

where Λ_n is the "approximate sensitivity matrix" (as in Andrews et al. (2017)) of parameters to moments, and $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$ is the moment approximation error. Once the model has been estimated using f_n, this formula incurs minimal computational cost, so it is easy for the researcher to estimate the approximation error.

The choice of the parametric function f_n is crucial for achieving a good approximation. We provide convergence speed results for kernel smoothers and neural nets. The kernel smoother converges at least as fast as $n^{-1/K}$, where K is the dimension of θ. Using classic results from the literature, we find that neural nets converge at least as fast as $\left(\frac{\log n}{n}\right)^{1/4}$. For complex models (where K is large), neural nets converge faster. This intuition is classical in the machine-learning literature: Neural nets can "mitigate the curse of dimensionality" by increasing in complexity as the training sample size grows (see e.g. Barron (1994), and more recently Farrell et al. (2021)).

While the continuity of moments f is necessary for convergence, the structural model generating the moments can be discontinuous. The intuition is that, as long as, for given θ, model discontinuities have measure zero, moments of the model's variables will be generically continuous in θ (see Newey and McFadden, 1994, Lemma 2.4). For instance, this result applies to models of corporate investment with fixed adjustment costs. In these models, investment "jumps" around a low and a high threshold of shadow price (Abel and Eberly, 1994), but is otherwise continuous. This is enough to ensure continuity of the moments of these models. Overall, all "well-defined" economic models (in the sense that their discontinuities are defined by boundaries over variables) will generate continuous moments. We make this argument more precisely in Section 3.5.

We apply our methodology to two standard structural models in finance: (1) a dynamic corporate finance model similar to Hennessy and Whited (2007) (Section 4), and (2) a life-cycle consumption and portfolio choice model similar to Viceira (2001) and Cocco et al. (2005) (Section 5). In both case, we use neural networks to fit the approximate moment function f_n, since they exhibit faster convergence in high-dimensional settings.

We assess the "out-of-sample" performance of the approximate parameter estimates by drawing a separate evaluation set of parameters and corresponding moments. By targeting the moments in this set, we estimate the model with the approximation f_n and compare the resulting parameters to the true parameters that generated the out-of-sample evaluation set. In both applications, the correlation between the true parameters and our "approximate" estimates in the evaluation set is always larger than .95, and in many cases larger than .99. While our approach is orders of magnitude faster than SMM (about 1 second compared to an hour), the difference in estimates is small, if not negligible.

The approximate moment function f_n allows researchers to conduct various robustness checks. We show how this can be done in our two applications. First, we examine the sensitivity of parameter estimates to the choice of targeted moments. In structural works, discussions of identification typically focus on how the function f varies near the estimate θ^*. However, as noted by Andrews et al. (2017), a more effective diagnostic tool is the assessment of how parameter estimates vary with empirical moments, since it reveals how changes in data affect estimates. This approach is rarely employed due to the computational expense of reestimating the model while slowly adjusting empirical moments. Our method allows for this analysis at nearly no cost. In our applications, representing parameters as functions of moments offers a different, and often more accurate, diagnostic for identification.

Second, we evaluate the robustness of parameter estimates to the *selection* of targeted moments. For any structural model, the number of moments one can use to estimate the model is potentially large. If the model correctly represents the data-generating process, the specific moments targeted in estimation do not matter, as long as they identify all parameters. However, in practice, models are misspecified, so that moment selection matters. To evaluate robustness to moment selection, researchers usually calculate additional moments in the model not used in estimation and compare them to their empirical values. This approach is not ideal, as the choice of baseline moments is arbitrary. Moreover, if the estimated model fails to match these non-targeted moments, it is unclear whether and how matching these moments would affect parameter estimates. A better approach is to reestimate the model across many sets of possible moments. While this approach would be computationally expensive using traditional methods, this becomes feasible using our approximation f_n. In our two applications, we explore thousands of possible moment combinations, and report

the resulting distribution of parameter estimates. We find that few estimates are robust to moment selection (in the corporate finance model more so than the life-cycle model), and isolate the moments that significantly impact estimation.

Third, we test the robustness of estimates across different subsamples. In reduced-form analyses, econometricians often reestimate regression models on various subsamples to check whether estimates are stable or change in expected directions. In structural work, the equivalent exercise would require reestimating the model for each subsample, which becomes computationally prohibitive if there are many sample splits. However, the low computational cost of our approach makes these checks straightforward. As demonstrated in our corporate finance application, structural parameters exhibit considerable variation over time, providing valuable insights into the model's validity.

Finally, our methodology allows us to assess model misspecification. Specifically, we examine how inference is affected when the data is generated using alternative models, i.e., not the model used in the baseline estimation. We first simulate many datasets using alternative models with different specifications and a range of parameter values. For each of these datasets, we then reestimate structural parameters using the baseline model. Despite the many structural estimations required, this approach is feasible thanks to our approximation method. More formally, we specify an alternative model with structural parameters θ_a and moments $g_a(\theta_a)$. For each θ_a, we estimate the parameters of the baseline model that best fit the alternative model-generated moments by solving:

$$\hat{\theta}_n(\theta_a, g_a) = \arg\min_{\theta} \left(g_a(\theta_a) - f_n(\theta)\right)^\top W \left(g_a(\theta_a) - f_n(\theta)\right),$$

This allows us to estimate parameters for a wide range of θ_a values and alternative models g_a. For θ_a in the neighborhood of θ^* and g_a in the neighborhood of f, this method aligns with the diagnostic proposed by Andrews et al. (2017), which does not require specifying g_a. Our approach enables a broader exploration of potential misspecification, although it necessitates specifying the nature of misspecification (i.e., to take a stand about the alternative model g_a).

We apply this approach to our corporate finance model, considering alternative models where financing may be constrained by cash-flow expectations (Lian and Ma, 2021). We find that, when the baseline model ignores these cash-flow constraints (is

misspecified), the cost of financing constraints is overestimated by a factor of 2 to 3.

Layout. Section 2 provides a short literature review. Section 3 describes our approach and establishes the convergence results and approximation error. Section 4 applies the approach to a dynamic corporate finance model, and Section 5 applies it to a life-cycle portfolio choice model. Section 6 concludes.

2 Related literature

Our paper contributes to the extensive literature that uses the method of simulated moments to structurally estimate economic models. We demonstrate the advantages of our approach in the context of dynamic models of financial constraints (e.g. Hennessy and Whited, 2007; Midrigan and Xu, 2014); and portfolio choice (see Gomes, 2020, for a survey). However, these broad model classes also appear widely across other fields, including labor economics (e.g. Guvenen and Smith, 2014; Berger et al., 2025); banking (e.g. Begenau, 2020), housing markets (e.g. Greenwald and Guren, 2021; Piazzesi et al., 2020), and household finance (e.g. Ameriks et al., 2020).

Our primary contribution aims at increasing transparency in the structural estimation of these models, with a particular focus on the sensitivity of policy predictions to moment or model misspecification. Andrews et al. (2020b) offers a formal definition of transparency in empirical research and applies it to structural estimation in economics. Andrews et al. (2017) derives a local linear approximation of the relationship between parameter estimates and data moments, providing a diagnostic tool for assessing misspecification bias. This method does not require explicitly specifying the misspecification but assumes that misspecification bias is small. While our analysis of misspecification is global, it does require to specify alternative models to the baseline model. In a follow-up work, Andrews et al. (2020a) formalizes the link between descriptive analysis and structural estimation using a similar local approximation. Our approach facilitates exploring additional descriptive statistics at low cost (non-targeted moments, model outputs). More broadly, our work connects to the literature on robustness to model misspecification (e.g. Huber, 2011; Armstrong and Kolesár, 2021; Bonhomme and Weidner, 2018).

In the structural literature, discussions of identification typically revolve around the relationship between moments and structural parameters – the function $f(\theta)$.

Table D.1 reviews recent structural papers in corporate and household finance. Of the 43 papers surveyed, 12 show local comparative statics (e.g., plots of f around θ^*), 8 report the Jacobian matrix (derivatives of f around θ^*), and 24 omit both. Four recent papers reports the sensitivity matrix of Andrews et al. (2017), i.e. the local derivative of parameter estimates w.r.t. targeted moments. Our approach eliminates the need for linear approximations, which, as our examples show, are not always valid.

Our paper is also related to a growing literature that seeks to improve numerical solutions for *models* through approximations. For instance, Norets (2012) uses neural networks to approximate the solution of a finite-horizon, dynamic discrete choice model. Similarly, Duarte (2020) describes a new solution method combining ML algorithms and Gradient Descent Algorithm. Chen et al. (2021) and Fonseca et al. (2022) use predictive algorithms to estimate the solution of models, after these are numerically solved on a training sample.[2] The key idea in our approach is to approximate *moment functions*, rather than *value or policy functions* of the underlying model. Moment are usually well-behaved functions of parameters, even if the underlying model features highly non-linear or discontinuous policy functions. Our approach is driven by our interest in estimation and robustness analyses, rather than solving the model.

Like us, Creel et al. (2015) and Creel (2021) do not approximate the model, but they take the approach opposite to ours. While our paper approximates the function $f(\theta)$ linking parameters to moments, their method directly approximates the inverted mapping from data moments m to parameters θ. Such inverted mapping may not, in practice, be a well-defined function, for instance when the model is not identified or rejected by the data. Further, the inverted mapping from m to θ is specific to a given set of target moments and a given weight matrix—therefore, it cannot be used to reestimate the model in alternative samples or using different moments. By approximating the moment function, our method is more generall and allows to conduct robustness checks more broadly.

Finally, an established literature outside economics, called surrogate optimization approximates the loss function to speed up a global optimization routine (see variations of this approach in Jones, 2001). In the context of structural estimation, the SMM loss function depends also on the targeted data moments, which varies

[2]For further references, see Fernández-Villaverde et al. (2021), Villa and Valaitis (2019), Maliar et al. (2019), Azinovic et al. (2019).

across robustness tests. With our approach of approximating simulated moments as a function of parameters, and then using the approximated moments in the SMM loss function, we are able to provide extensive robustness analyses with varying data moments and weight matrix in the estimation applications.

3 Our approach: Presentation and general results

This section outlines our approach and the conditions under which it applies. We show convergence, and provide a formula to compute the approximation error. We then supply results on convergence speed. Finally, we explain why our approach does *not* require that the data-generating model is itself continuous. It only requires that moments are continuous, a much milder assumption.

3.1 Description and notations

Let $\mathcal{S}$ be a structural model with deep parameters $\theta \in \mathbb{R}^K$, generating a vector of moments $f(\theta) \in \mathbb{R}^M$, where M is the number of computable moments. These moments do not need to be all empirically observable. Some can be a combination of structural parameters that capture an outcome of interest (e.g., the average loss in market value due to financial constraints in a structural corporate finance model).

In simulation-based estimation, $f(\cdot)$ does not admit a closed-form representation and is computed numerically. For clarity, we assume that $f(\cdot)$ can be exactly computed via extensive simulations, ignoring any simulation error. Let m represent the vector of empirical counterparts to the model-based moments $f(\theta)$. The minimum distance estimator of θ is obtained by the following minimization:

$$\theta^* \in \arg\min_{\theta} \left(m - f(\theta)\right)^\top W \left(m - f(\theta)\right), \tag{1}$$

where W is a weighting matrix, which assigns zero weights to moments that cannot be observed in the data. We call θ^* the *true* parameter estimate, i.e., the solution to Equation (1).

Estimating the model for a specific set of moment values, m, is computationally expensive. When the model lacks a closed-form solution, it must be solved numerically and moments have to be generated through simulation. This process is

time-consuming, especially for dynamic and non-linear models. Additionally, optimization algorithms must compute moments $f(\theta)$ across many parameter values to find θ^*, the global minimizer in Equation (1). This issue is exacerbated by the curse of dimensionality, as the parameter space expands exponentially with K.

Due to these limitations, researchers often perform only a few estimations. Fundamentally, the issue that plagues this approach is that it lacks "economies of scale": each new estimation on a different set of empirical moments (whether on moments estimated for different subsamples or a different set of moments altogether) incurs the same computational cost as the initial estimation.

Our approach creates such economies of scale using an approximation of the function $f(\cdot)$. To achieve this, we propose the following process:

1. **Parameter Bounds:** Define ex ante bounds for each parameter in θ. These bounds also need to be specified in standard estimation techniques. They can be based on expert knowledge or prior findings in the literature. Let $\mathbf{P}_\theta \subset \mathbb{R}^K$ be the set of admissible parameter vectors.

2. **Drawing Parameters:** We select n parameter values $(\theta_i)_{1\leq i\leq n}$ from this set $\mathbf{P}_\theta$. The exact method to build this estimate is not critical. One option is to use a regularly spaced grid. In our applications, we use quasi-random Halton sequences, which maintain uniform density as new points are added. Thus, to increase the size of the training sample from n to n', the researcher only needs to draw $n' - n$ parameter values while retaining the original n draws.

3. **Model Simulations:** For each parameter values θ_i, we solve the model and simulate moments $f(\theta_i)$. This step creates a *training dataset* $\mathcal{D}_n$, consisting of n parameters and their corresponding moments under the model $\mathcal{S}$. Although computationally intensive, this step is performed only once.

4. **Fitting the Approximation:** We use the training data $\mathcal{D}_n$ to find a parametric approximation $f_n(\theta)$ of $f(\theta)$, which can be computed for any θ. We refer to f_n as the *approximate moment function*. While no specific parametrization is required, we focus on neural networks, as they converge faster than kernel methods and yield more accurate results in our applications.

5. **Estimation:** Estimate $\hat{\theta}_n$ as the solution of:

$$\hat{\theta}_n \in \underset{\theta \in \mathbf{P}_\theta}{\arg\min} \, (m - f_n(\theta))^\top W (m - f_n(\theta)). \tag{2}$$

where $\hat{\theta}_n$ is the *approximate* parameter estimate.

Only step 3, building the training dataset $\mathcal{D}_n$, is computationally expensive. However, the size of the training dataset is typically chosen to match the number of model simulations required for a single estimation using SMM. Therefore, this step is no more costly than a standard estimation.

3.2 Convergence

Let m be a vector of empirical moments. To simplify the exposition, we assume that m is measured without statistical error, since this aspect of the inference problem is already covered by well-known formulas (see, e.g., Gouriéroux and Montfort, 1996). We make the following assumptions:

Assumption 1. *Suppose that the following hold:*

i. $\mathbf{P}_\theta$ *is nonempty and compact.*

ii. W *is positive-definite.*

iii. There exists a unique $\theta^* \in \mathbf{P}_\theta$ *such that* $m = f(\theta^*)$.

iv. The model $\mathcal{S}$ *generates a continuous moment function* $f(\cdot)$.

The first two assumptions are standard and technical. The third assumption is stronger, as it requires that the model is (1) well-identified and (2) fits the data, though we do not assume that the model is the true data-generating process. The fourth assumption requires moments to be *continuous* functions of parameters. This continuity assumption is crucial but is satisfied in all well-defined economic models, including the ones that feature discontinuities (like investment models with fixed adjustment costs for instance). We explain in detail why in Section 3.5.

The following theorem, a non-random adaptation of Newey and McFadden (1994, Theorem 2.1) to our setting, formalizes the conditions under which the approximate parameter estimate (the minimizer of program 2) converges to the true parameter estimate (the minimizer of program 1):

Theorem 1 (Convergence of approximate parameter estimate)**.** *Suppose Assumption 1 holds and the approximate moment function f_n satisfies:*

i. f_n uniformly converges to f as $n \to \infty$.

ii. f_n is bounded.

Then, the approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$ as $n \to \infty$.

Proof. See Appendix A.2. □

The conditions (i) and (ii) in Theorem 1, requiring that f_n converges uniformly and is bounded, can be ensured by the choice of the approximation f_n.

3.3 Asymptotic approximation error formula

When n is large but finite, our approximate estimator differs from the actual estimator since $f_n \neq f$. This difference can be computed analytically provided n is sufficiently large, i.e., when $\hat{\theta}_n - \theta^*$ is small enough:

Proposition 1. *Suppose the assumptions in Theorem 1 hold. Additionally, assume there exist N_0, L_0, and a neighborhood $\mathbf{O} \subseteq \mathbf{P}_\theta$ around θ^* such that $\mathbf{O}$ is open in $\mathbb{R}^K$ and for $n \geq N_0$, f_n is continuously differentiable on $\mathbf{O}$. Moreover, f is twice differentiable with a bounded second derivative on $\mathbf{O}$. Moreover, the sensitivity matrix Λ_n, defined below, exists with $\|\Lambda_n\|_2 \leq L_0$. Then, as $n \to \infty$, we have:*

$$\hat{\theta}_n - \theta^* = \underbrace{\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)\right)^{-1} \nabla f_n(\hat{\theta}_n)^\top W}_{\equiv \Lambda_n} \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right) + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right). \tag{3}$$

Proof. See Appendix A.3. □

A key additional assumption in Proposition 1 is that f is locally twice differentiable around the true parameter. The assumption that f_n is continuously differentiable is easier to satisfy if f_n is constructed appropriately (e.g., using a neural network).

Another crucial assumption in Proposition 1 is the existence and boundedness of the matrix Λ_n. Intuitively, Λ_n is similar to the "sensitivity matrix" in Andrews et al. (2020b): it represents the local derivative of parameters with respect to moments.

The existence of Λ_n requires that the model is identified (i.e., ∇f is full-rank), along with some regularity conditions (since the matrix inversion also involves ∇f_n). For simplicity of exposition, we directly require the boundedness of $\|\Lambda_n\|_2$ rather than imposing more primitive conditions. Finally, the second term on the right side of Equation (3), $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$, measures the approximation error on the moments themselves (at the approximated parameter estimate).

The formula introduced in (3) is computationally inexpensive to obtain. While the values of $f(\hat{\theta}_n)$ and its gradient $\nabla f(\hat{\theta}_n)$ must be computed numerically, this requires only a small number of evaluations. The functions f_n and their gradients at the estimate are simple enough to often have closed-form expression.

In our applications in the next sections, we demonstrate that the formula (3) provides an accurate approximation of the true error.

3.4 Two parameterizations for f_n and speeds of convergence

We now turn to the construction of f_n. We explore two parameterizations that ensure that the conditions in Theorem 1 are satisfied (in particular, that $f_n \to f$ uniformly), and for which we can produce results on convergence speed.

3.4.1 Granularity of the training data

We first introduce a key concept, the *granularity* of the training data. For the parameters in the training sample, $(\theta_i)_{1 \leq i \leq n}$, we define:

$$\delta_n = \max_{\theta \in \mathbf{P}_\theta} \min_{1 \leq i \leq n} \|\theta - \theta_i\|_2, \tag{4}$$

which represents the maximum distance from any point in $\mathbf{P}_\theta$ to its nearest element in the training dataset $\mathcal{D}_n$. Our asymptotic results assume that as n increases, $\delta_n \to 0$: each point on $\mathbf{P}_\theta$ becomes closer and closer to the elements of the training dataset, i.e., the training data become infinitely granular.

How fast does δ_n decrease with n, the size of the training dataset $\mathcal{D}_n$? The answer depends on how the elements of $\mathcal{D}_n$ are structured.

To build intuition, consider the simple case of a regular grid over $\mathbf{P}_\theta$. Suppose the set of possible parameter values is a cube of dimension K, and that $\mathcal{D}_n$ consists of regularly spaced points in this cube. Then, it is easy to see that:

$$\delta_n = \mathcal{O}\left(n^{-\frac{1}{K}}\right).$$

Thus, δ_n decreases with n, but more slowly if parameters have a large dimension K – the standard curse of dimensionality.[3]

In our application, we use a uniform draw over $\mathbf{P}_\theta$ via a Halton sequence instead of a regular grid. The advantage of this method is that granularity increases uniformly by simply adding more points to the existing sample, unlike a regular grid that would require redrawing the entire sample to increase n, which is computationally costly.

With uniform draws, the upper bound on convergence is the same as with a regular grid, although only in probability, as shown in the following lemma:

Lemma 1. *Suppose* $\mathbf{P}_\theta = [a_1, b_1] \times \ldots \times [a_K, b_K] \subset \mathbb{R}^K$, *with* $a_i < b_i < \infty$ *for* $1 \leq i \leq K$, *and assume* $\mathcal{D}_n$ *is generated by draws of* θ *from a uniform distribution over* $\mathbf{P}_\theta$. *Then, for any value of* $0 < p < 1$, *with probability* $1 - p$, *the maximum distance to the nearest training data point,* δ_n, *defined in* (4), *satisfies:*

$$\delta_n = (-\log p)^{\frac{1}{K}} \cdot \mathcal{O}\left(n^{-\frac{1}{K}}\right). \tag{5}$$

Proof. See Appendix A.4. □

This lemma shows that, even when requiring a high confidence (i.e., p close to 0), the upper bound on δ_n decreases nearly as fast as $n^{-1/K}$. Thus, the granularity of uniform random draws is, asymptotically, similar to regular grids.

3.4.2 Kernel approximation

Next, we examine the properties and convergence speed of kernel smoothing (Li and Racine, 2023), a natural parameterization for f_n. We use a Nadaraya-Watson weighted average formula, while requiring specific properties of the kernel function that best fit our setting. Given training data $(\theta_i)_{1 \leq i \leq n}$, the kernel smoothing function is defined as:

$$f_n(\theta) = \frac{\sum_{i=1}^n k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i=1}^n k_n(\theta, \theta_i)}, \qquad k_n(\theta, \theta') = \eta\left(\frac{\|\theta - \theta'\|_2^2}{\lambda_n^2}\right), \tag{6}$$

[3] To see this, assume edges of $\mathbf{P}_\theta$ have length E_k and each edge has q regularly spaced points. Then, the number of points in the training sample is $n = q^K$. The maximum distance between the grid and any point of $\mathbf{P}_\theta$ is $\delta_n = \frac{\sqrt{\sum_k E_k^2}}{2q}$. Thus, $\delta_n \propto n^{-\frac{1}{K}}$.

where the weight function $\eta(\cdot)$ is a real-valued, continuously differentiable function that accepts scalar inputs. Moreover, $\eta(x) > 0$ for $0 \leq x < 1$, $\eta(x) = 0$ for $x \geq 1$, and $\eta(x)$ is non-increasing for $0 \leq x \leq 1$. The scaling parameter λ_n is set to be larger than the granularity of the training dataset δ_n, i.e., $\lambda_n = \gamma\,\delta_n$ for some constant $\gamma > 1$.

The following proposition summarizes the properties of kernel smoothing as $\delta_n \to 0$:

Proposition 2. *Let $\hat{\theta}_n$ be the approximate parameter estimate based on the kernel-smoothing function in (6). Suppose Assumption 1 holds. Then, as $n \to \infty$ and the training data become more granular ($\delta_n \to 0$), we have:*

i. The approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$.

ii. Furthermore, ∇f_n exists and is continuous on any neighborhood $\mathbf{O} \subseteq \mathbf{P}_\theta$ around θ^* that is open in $\mathbb{R}^K$. Assuming the rest of assumptions in Proposition 1 hold and $\|\nabla f(\theta^*)\|_2 \neq 0$, the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(\delta_n\right). \tag{7}$$

iii. Assuming the conditions in part (ii) hold, if the training data lie on a regular grid, the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(n^{-\frac{1}{K}}\right). \tag{8}$$

Finally, as shown in Lemma 1, the same bound holds in probability when the training data is uniformly drawn from a bounded box in $\mathbb{R}^K$.

Proof. See Appendix A.5. □

Proposition 2 shows that, under regularity conditions, the convergence speed is at least of the same order as δ_n, the granularity of the training data. Additionally, if the training data are evenly spaced or drawn uniformly, the convergence error is given by $n^{-1/K}$. This is the classic curse of dimensionality: a large number of parameters dramatically slows down convergence speed.

3.4.3 Neural network approximation

We now explore the properties and convergence speed of approximate moments when f_n represents a neural network. To leverage existing results from the literature, we focus here on a training dataset generated by random draws of θ in $\mathbf{P}_\theta$. As we show below, neural networks converge faster than kernel, especially for high-dimensional parameters.

Assume each component function f_n^r, $1 \leq r \leq M$ of f_n is a shallow neural network (NN) of the class:

$$f_n^r(\theta) = c_0^r + \sum_{i=1}^{N} c_i^r \phi\left((a_i^r)^\top \theta + b_i^r\right), \tag{9}$$

where N is the number of nodes, ϕ is a sigmoid function which is infinitely differentiable, a_i^r is a K-dimensional vector, while c_0^r, c_i^r, and b_i^r are scalars.

The exact convergence speed of f_n depends on several factors, including the loss function used to train the network, the distribution of θ, or the smoothness of the true function f. For simplicity, we rely here implicitly on the assumptions needed to show convergence and convergence speed of f_n in Barron (1994). These assumptions yield the following result for the approximate parameter estimate $\hat{\theta}_n$, labeled as "informal" since we do not fully specify the assumptions:

Proposition 3 (Informal). *Let $\hat{\theta}_n$ be the approximate parameter estimate based on a neural network described in (9), which is appropriately trained over the training dataset $\mathcal{D}_n$, generated by random draws of θ in $\mathbf{P}_\theta$. Moreover, suppose Assumption 1 holds. Then, as the training data size $n \to \infty$:*

i. The approximate parameter estimate converges to the true parameter estimate: $\hat{\theta}_n \to \theta^*$.

ii. Suppose the neural net training procedure is appropriately set up such that f_n, together with f and θ^*, satisfies the assumptions of Proposition 1 and $\|\nabla f_n(\theta)\|_2$ is uniformly bounded across n and θ. Moreover, let the number of neural net nodes increase as $N = \sqrt{n/(K \log n)}$. Then, the convergence error satisfies:

$$\|\hat{\theta}_n - \theta^*\|_2 = \mathcal{O}\left(\sqrt{M}\left(\frac{K \log n}{n}\right)^{\frac{1}{4}}\right). \tag{10}$$

Proof. See Appendix A.6. □

This informal proposition leverages both our earlier results and two important findings from the early machine-learning literature. First, Cybenko (1989) shows that neural networks uniformly converge toward continuous functions, which allows us to apply Theorem 1, establishing that $\hat{\theta}_n \to \theta^*$. This property, along with some regularity conditions, enables us to apply Proposition 1, which relates $\hat{\theta}_n - \theta^*$ to $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$. Finally, we use Barron (1994) (and its associated regularity conditions) to establish an upper bound on the convergence of $f(\hat{\theta}_n) - f_n(\hat{\theta}_n)$, which leads to our final result.

Proposition 3 shows that the convergence speed of the neural network approximation depends less on the dimension K compared to the kernel approximation, where K enters multiplicatively but not as a power of n. This result echoes standard findings in machine learning. Neural networks, by increasing the number of nodes, can better handle complexity and mitigate the curse of dimensionality.

While this result holds for a single layer NN, it is common to use, in practice, deeper NNs, rather than the shallow but very wide structure suggested by Barron (1994). At the very least, the above result provides an upper bound for the convergence error of deeper neural networks (i.e., a lower bound for the convergence speed). See, for instance, Farrell et al. (2021) for recent convergence speed results in the case of Multi-Layer Perceptrons (MLPs). Using such advanced results seems promising, but beyond the scope of this paper.

3.5 What ensures the continuity of moments

A key assumption to ensure the convergence of our method is the continuity of the true moments $f(\theta)$. Although this assumption may appear restrictive, *it does not require the model itself to be continuous*. The only condition is that, for given parameter value θ, the discontinuities of the moment-generating function have measure zero. This condition is satisfied in all well-specified economic models, even those with discontinuities, as we discuss below.

Consider a generic model parametrized by $\theta \in \mathbb{R}^k$, with variables $x \in \mathbb{R}^d$ (some endogenous, some exogenous). For example, in the corporate finance aplication of Section 4, x may include investment, debt, and productivity shocks, while θ contains deep parameters such as adjustment costs and productivity volatility. Further, let $h(x, \theta)$ be the vector of functions of x, whose expectation defines the moments of

interest, i.e. $f(\theta) = \mathbb{E}_x[h(x,\theta)]$. In this framework, Newey and McFadden (1994, Lemma 2.4) show that $f(\theta)$ is continuous if the discontinuity points of $h(x,\theta)$ have total probability zero along θ. For instance, this holds when the discontinuities of $h(x,\theta)$ occur at "points" along each θ under a continuous distribution.

To visualize the logic behind this result, consider the following simple example. Assume $x > 0$ is distributed according to p.d.f. f, and consider the function $h(x,\theta) = x\mathbf{1}_{\{x>\theta\}}$. Clearly, h is discontinuous in $x = \theta$ if $\theta > 0$. However, $f(\theta) = \mathbb{E}_x[h(x,\theta)] = \int_{x>\theta} xf(x)dx$ is not only continuous but differentiable with respect to θ. The same intuition holds for a higher number of breakpoints, provided they are countable.

The assumption that the set of discontinuities has measure zero is satistifed by all well-specified economic models. Take for instance the case of models of investment with non-convex adjustment costs. Abel and Eberly (1994) show that in the presence of fixed and linear adjustment costs, investment is discontinuous in two thresholds of shadow price: It is equal to zero between these two thresholds, positive above the high one, and negative below the low one. This is a case where discontinuities occur at "points" for given values of structural parameters. Newey and McFadden (1994, Lemma 2.4) ensures that moments of investment and other variables will be continuous w.r.t. the model's deep parameters.

4 Application to a dynamic corporate finance model

To illustrate the power of our approach, we apply it to two popular models in finance. We will first discuss the performance and speed of our method, and then use it to perform various robustness checks (to moment selection, misspecification, etc.). This section implements this on a standard corporate finance model. In Section 5, we do the same for a portfolio choice model along the life cycle.

4.1 Model, parameters and moments

We use a standard model of firm dynamics with collateral constraints, as described in Catherine et al. (2022). Time is discrete. Every period, the firm draws a productivity z_t, and given the current capital k_t and debt d_t chooses the next period's capital k_{t+1} and debt d_{t+1} to maximize a discounted sum of per-period cash flows, subject to financing constraints. The discount rate is $r = 3\%$.

At time t, the firm's profit is: $\pi_t = e^{z_t(1-\alpha)}k_t^{\alpha}$. Productivity follows an AR(1) process $z_{t+1} = \rho_z z_t + \eta_{t+1}$, where the variance of the innovation term η_{t+1} is σ_z^2. Investment $i_t = k_{t+1} - (1-\delta)k_t$ entails a convex cost $\frac{\gamma}{2}\frac{i_t^2}{k_t}$. δ is the depreciation rate and γ is the adjustment cost rate. Profits, net of interest payments and depreciation, are taxed at a rate $\tau = 1/3$. This tax rate applies both to negative and positive profits so that firms receive a tax credit when their accounting profits are negative. The firm can hold cash (when $d_t < 0$) or issue debt ($d_t > 0$), which is risk-free and incurs an interest rate r.

Financing frictions arise from two constraints: (1) a collateral constraint that limits borrowing $d_{t+1} \leq \lambda k_{t+1}$, and (2) a linear cost (ξ) of equity issuance, such that negative cash-flows to equity are multiplied by $(1+\xi)$.

The Bellman equation for the firm's optimization problem is written as

$$\begin{aligned} V(k_t, d_t; z_t) = &\max_{k_{t+1}, d_{t+1}} CF + \frac{1}{1+r}\mathbb{E}[V(k_{t+1}, d_{t+1}; z_{t+1})|z_t], \\ &\text{with: } CF = G\left(\pi_t - i_t - \tfrac{\gamma}{2}\tfrac{i_t^2}{k_t} + d_{t+1} - (1+r)d_t - \tau(\pi_t - r d_t - \delta k_t)\right), \\ &\text{such that: } d_{t+1} \leq \lambda k_{t+1}, \\ &\qquad \forall x: \; G(x) := x(1 + \xi \mathbf{1}_{x>0}). \end{aligned} \tag{11}$$

Moments and estimated parameters In total, there are seven structural parameters: $\theta = (\delta, \gamma, \alpha, \rho_z, \sigma_z, \xi, \lambda)$. We also compute the "value loss of financing constraints", which represents the difference in the mean log value of constrained firms compared to similar unconstrained firms (i.e., $\xi = 0$).

The literature has used a variety of moments to identify these parameters. We define f as the function linking the parameters θ to a set of 17 moments used in previous works. The first seven moments are: mean(investment/assets), mean(profit/assets), mean(equity issuance/assets), mean(leverage), mean autocorrelation of investment, std(log growth sales) and std(log growth 5yr sales). They correspond to the moments used in Catherine et al. (2022), who also discuss how they identify the model's parameters. An additional ten moments, used in the literature, are also constructed and used in our robustness analysis. Appendix B.1 describes these moments. Throughout our analysis, the dataset we use is a COMPUSTAT extract from 1971–2019.

4.2 Training the approximate moment function f_n

We now construct the approximate moment function f_n. We restrict structural parameters to a compact box $\mathcal{P}$ with the following ranges: $\delta \in [0; .2]$, $\gamma \in [0; .3]$, $\alpha \in [.5; .9]$, $\rho_z \in [.5; .98]$, $\sigma_z \in [.2; 2]$, $\xi \in [0; .3]$ and $\lambda \in [0; .6]$. We generate a training sample $\mathcal{D}_n$ using a Halton sequence of $n = 50,000$ parameters drawn from $\mathcal{P}$. For each draw, we solve the model, and calculate the 17 corresponding moments using simulations. This step takes approximately 140 hours with our numerical setup.

We also generate a separate out-of-sample evaluataion dataset of 1,000 parameters using a Halton sequence in $\mathcal{P}$. Some draws leads to models that cannot be identified. For instance, if the cost of equity issuance is sufficiently large, firms won't issue any equity. As a result, the ratio of equity issuance to assets cannot identify the cost of equity issuance beyond a threshold. This non-identification is not specific to our approximation, i.e., the true model is also not identified for these draws.

To detect these "badly identified" draws, we compute $(J^\top W J)^{-1/2}$ for each draw i in the evaluation set, where $J = \nabla f(\theta_i)$ is the Jacobian matrix of the true model and W is the inverse of the variance-covariance matrix, estimated by bootstrapping the full sample, for the seven baseline moments described in Section 4.1. We then drop draws for which any diagonal elements of this matrix is 10 times larger than the standard error of the full-sample parameter estimates (obtained through real SMM, see below).[4] We end up with a final evaluation set of 215 observations.

We train a neural network f_n using the training sample. In Appendix B.4, we compare this neural network approximation to a kernel method and show that the neural network achieves a better fit, consistent with the faster convergence of neural networks in high-dimensional problems (see Proposition 3). The neural network is a Multi-Layer Perceptron (MLP) with 5 layers and 512,256,128,64,32 nodes. Implementation details are in Appendix B.3. This architecture is more complex than in proposition 3, as deep neural nets have been shown to converge faster when the target function f is smooth enough (Farrell et al., 2021), especially for larger dimensions.

[4]While this selection criterion is somewhat arbitrary, we have experimented with alternative definitions and found that this did not affect our assessment of the estimates' precision.

4.3 Approximate moment estimation: Performance

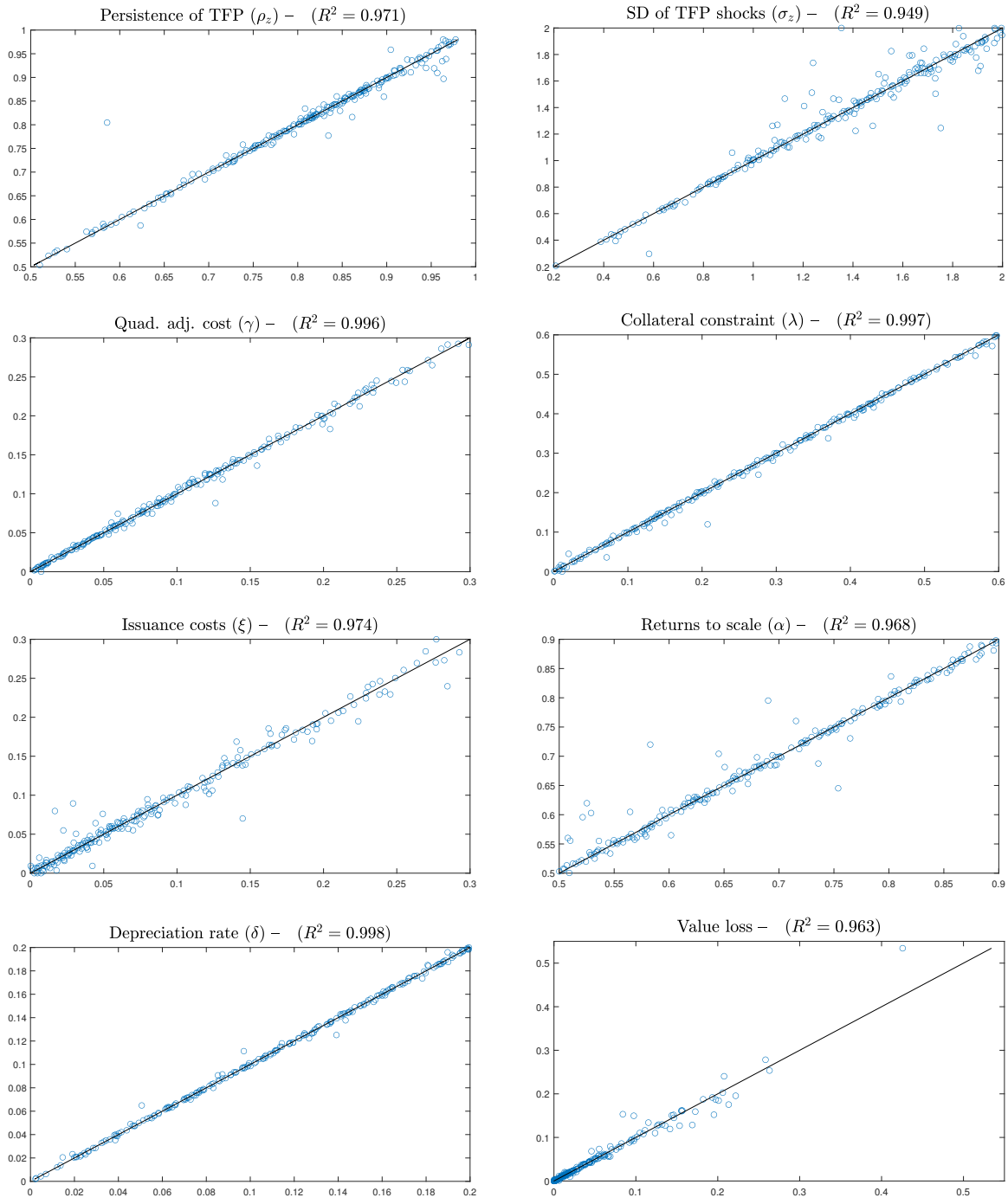
To measure the performance of our approach, we use two methods: (1) comparing the approximate estimates to the true parameters using the evaluation set and (2) comparing the approximate estimates to actual SMM estimates on real data.

First, for each draw of parameters and corresponding moments in the evaluation set, we target the moments and estimate the seven structural parameters using the *approximate* moment function f_n. We target the first seven baseline moments as described in Section 4.1. Figure 1 shows scatter plots comparing the estimated and true parameters, across all the draws in the evaluation set. R^2 values are above 95% for all parameters. We also observe that some parameters are better estimated than others. This is because, despite the filter applied to the evaluation set, some draws still lead to poorly identified models. This is in particular true for equity issuance costs, as low equity issuance is consistent with a large range of equity issuance cost parameters. Nevertheless, Figure 1 establishes that the method overall performs well.

Next, we compare the approximate estimates to actual SMM estimates on moments from real data. The advantage of this comparison is that we know that the model is well-identified when matched on real empirical moments. We conduct the SMM estimation using a standard procedure (TikTak algorithm): (1) we initialize the algorithm by evaluating the SMM objective at 50,000 different starting points (2) we run Nelder-Mead optimizations at the 50 best starting points using at most 200 function evaluations. The model is estimated by targeting the seven baseline moments in the full COMPUSTAT sample, shown in column 1 of Appendix Table B.1. Table 1 reports true and approximate SMM estimates, while Appendix Table B.1 provide moment fits. Table 1 shows that true and approximate estimates are all within a few percentage points of each other. Line 3 applies the approximation error correction from Section 3.3, which further reduce the error in the approximate estimates.[5]

[5] Appendix Figure B.2 shows that our approach is faster than the standard SMM by several orders of magnitude. The computing times we report exclude the simulation of the training sample, which is long but required for both estimations. For the true SMM, it takes an additional 20 minutes for the estimation to converge to its final value. In contrast, the approximation-based estimation converges in less than a second (provided the neural net has been estimated).

Figure 1: Out-of-sample performance (Corporate Finance Model)



Notes. This figure shows the precision, for the corporate finance model, the presision in the out-of-sample evaluation set of our benchmark approximate SMM across estimated parameters. For each draw $(\theta, f(\theta))$ in the evaluation set, we use neural nets that construct the approximate moment function f_n and estimate parameters $\widehat{\theta}_n$ which match $f(\theta)$. The x-axis reports the true parameters θ, and the y-axis reports the estimated parameters $\widehat{\theta}_n$.

Table 1: Parameter estimates, true vs approximate SMM (Corporate Finance Model)

	ρ_z	σ_z	γ	λ	ξ	α	δ	value loss
true SMM	.7178	1.1566	.0413	.1087	.0378	.8155	.0670	.0250
- s.e., local deriv.	.0068	.0508	.0024	.0030	.0021	.0074	.0008	.0019
approx. SMM	.7270	1.1413	.0439	.1070	.0378	.8141	.0669	.0254
- s.e., local fit deriv.	.0070	.0560	.0024	.0029	.0017	.0082	.0008	.0018
approx. SMM, corrected	.7179	1.1569	.0412	.1087	.0381	.8158	.0671	.0252
- s.e., local fit deriv.	.0065	.0525	.0022	.0030	.0016	.0077	.0008	.0017

Notes. The table reports the parameter estimates of the corporate finance model presented in Section 4.1. We report parameter estimates using the true SMM (first line), the approximate SMM (second line), and the approximate SMM corrected with the first-order error formula from Section 3.3 (third line). The approximate SMM uses neural net fit to approximate moment functions. 's.e., local deriv.' shows the standard errors of the true SMM parameters, calculated using the Jacobian matrix of the true model f. 's.e., local fit deriv.' shows the standard errors of the approximate SMM parameters, calculated using the Jacobian matrix of the approximate moment function f_n. We use the delta method to calculate the standard error of value loss.

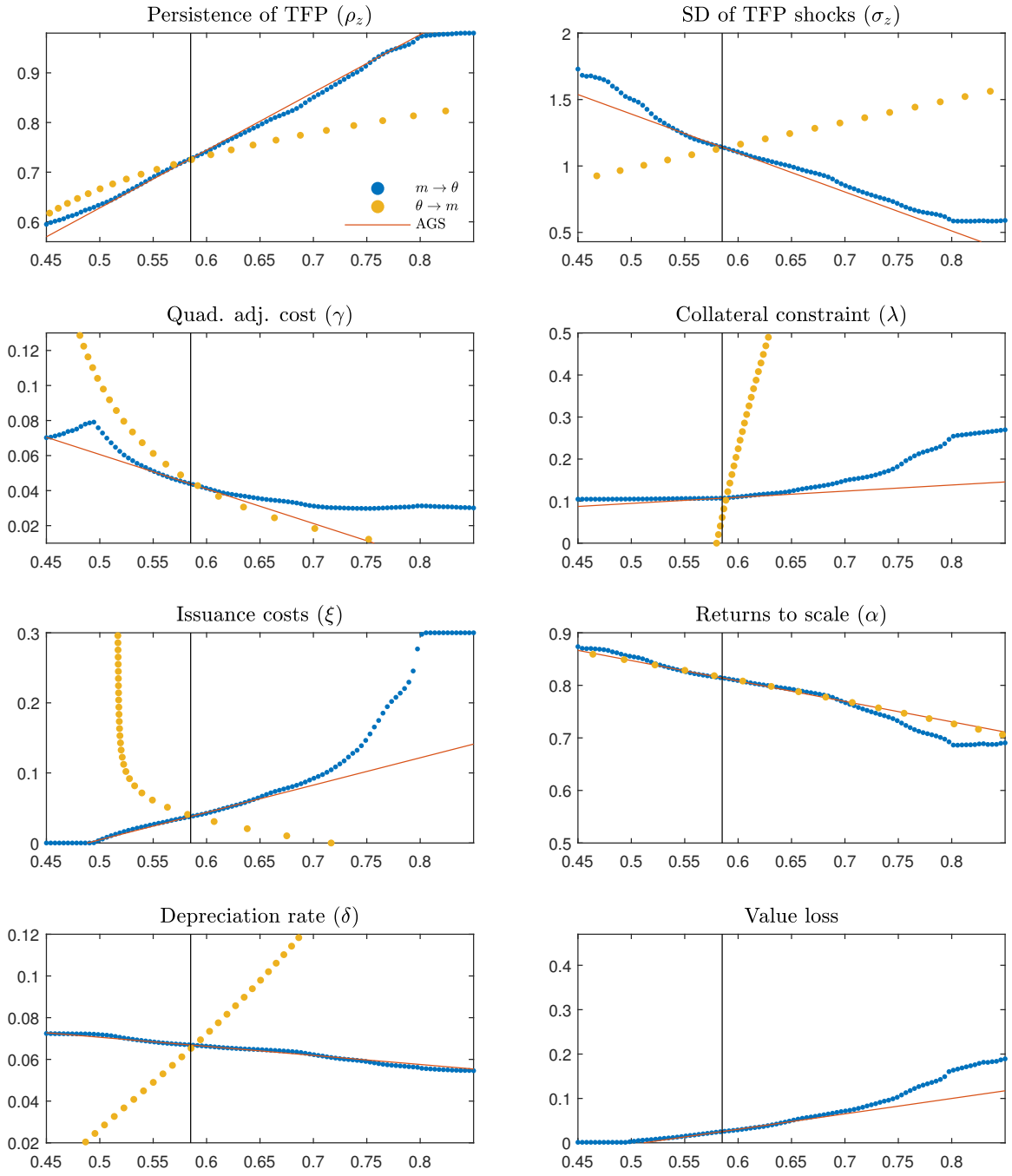
4.4 Identification diagnostic

Our method can assess how moments affect parameter estimates, a computationally intensive task requiring reestimating the model for many different values of the targeted moments.

For brevity, Figure 2 illustrates this approach by focusing on a specific moment, the volatility of 5-year sales growth. The yellow line shows how this simulated moment changes as a function of the different parameters. This is the typical "comparative static" figure reported in structural works. In contrast, the blue line corresponds to our identification diagnostic: we reestimate the model many times by targeting many possible values for the volatility of 5-year sales growth and report the relation between the approximate parameter estimates and the targeted moment values. Every point on the blue line thus corresponds to a different estimation. The red line shows the local linear approximation from Andrews et al. (2017), which corresponds to a linearized version of our technique around the SMM estimate and does not require reestimating the model.

Figure 2 illustrate the advantages of our methodology. First, the standard "comparative static" figure of the literature can be misleading. For instance, the yellow line on the top-right panel shows that a larger volatility of TFP shocks (σ_z) increases the volatility of 5-year sales growth. One could use this analysis to conclude that samples with larger volatility of 5-year sales growth would lead to infer a higher σ_z.

Figure 2: Sensitivity of parameters to moments: 5-year sales growth volatility (Corporate Finance Model)



Notes. This figure plots parameter values on the y-axis and the volatility of 5-year sales growth on the x-axis. The yellow line draws local comparative statics, i.e. how variations in one parameter around its estimated value – holding other parameters fixed at their estimated value – affect the moment value. The blue line plots how variations in the moment value – holding other moments fixed at their empirical value – affect the estimation of each parameter. Each dot on the blue line corresponds to a separate estimation using our approximation technique. The red line is the local approximation of Andrews et al. (2017). The black vertical line shows the value of the moment in the data.

This conclusion would be incorrect. The blue line on the same panel shows that a higher volatility of 5-year sales growth instead leads to infer *a lower* σ_z. The reason for this discrepancy is that the "comparative static" approach ignores that parameter estimates jointly depend on all targeted moments. This is easily seen in our context. For a given volatility of 1-year sales growth, a larger volatility of 5-year sales growth implies that TFP shocks are more persistent (i.e., a higher ρ, as shown in the top-left panel). Because TFP persistence is higher, capital becomes more responsive to TFP shocks, which, in turn, increases the volatility of 1-year sales growth. To match the data, the model then needs to *reduces* σ, the volatility of TFP shocks.

Second, Figure 2 also illustrates that the local linear approximation in Andrews et al. (2017) can fail to hold outside of a close vicinity of the actual moments used in the estimation. For instance, consider the case where the volatility of 5-year sales growth was 80%, instead of its empirical value of 58%. The linear approximation on the red line suggests that the collateral constraint parameter (λ) would be close to the one we estimate in the main estimation, leading to the conclusion that financial constraints would be about the same in this alternative sample. Our diagnostic shows that λ would be about twice larger, suggesting significantly lower financial constraints.

4.5 Robustness to moment selection

Methods of moments do not provide guidance on which moments to target in estimation. We can however use our approximation approach to evaluate the robustness of parameter estimates to moment selection. Conceptually, this is straightforward: we consider many combinations of possible targeted moments and reestimate the model for each combination. Robustness is established if parameter values remain stable across most estimations. This robustness exercise cannot be done with standard estimation techniques, as it requires many estimations. Our approximation approach makes this feasible at low computational cost.

Concretely, we reestimate our model by targeting both the 7 moments used in our baseline and any possible subset of the remaining 10 moments. This requires $2^{10} = 1024$ separate estimations. Of these, we only consider estimations that correspond to reasonably well-identified sets of moments: we drop cases where the standard errors for an estimated parameter is more than 10 times larger than the standard error of the baseline true SMM. This leaves us with 722 estimations. Figure 3 reports the

distribution of parameter estimates. The black lines show the baseline estimates, together with the 95% confidence interval. A baseline estimate would be robust to moment selection if a large share of alternative estimates are close to the baseline.

Most parameter estimates in this model are not robust to moment selection. For instance, the estimated variance of innovations to z, σ_z, which has a benchmark estimate of 1.1, frequently admits a wide range of alternative estimates – from 0.4 to 1.1 – depending on which moments are targeted in the estimation. We also see that in many alternative estimations, the parameter estimate for σ_z hits its upper bound of 2. Similar conclusions are obtained for all parameters.

4.6 Sample splits

Our method also allows for reestimating the model across subsamples, a common practice in reduced-form analysis but computationally expensive for structural models. Such sample splits can be useful as they can help test for specific mechanisms or identify useful trends in the data.

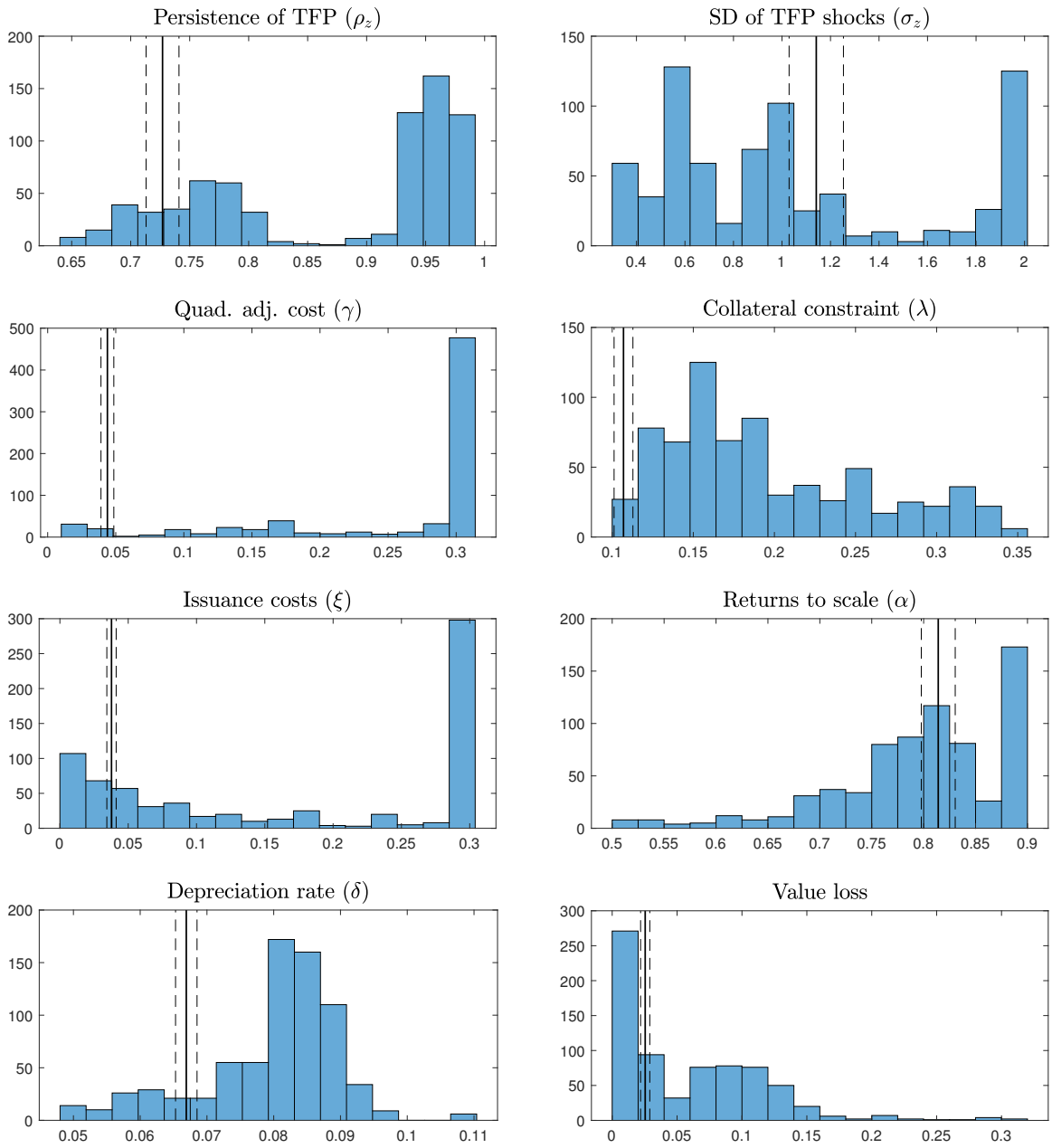
Figure 4 shows time-varying estimates of parameters: every year t, we reestimate the model by targeting moments estimated over the $[t-5; t+4]$ period in the data. This analysis reveals interesting trends. For instance, the collateral constraint parameter λ goes from .2 in the 1970s to close to 0 in the 2000s. One explanation is the well-documented increase in cash holdings over this period (Bates et al., 2009): a reduction in net leverage leads the model to believe that firms are more financially constrained and that the collateral parameter λ is thus smaller. Similarly, the estimated depreciation rate steadily declines from 8% to 4%. One possible explanation is the rise in intangible capital over the period (Crouzet and Eberly, 2018). As the model does not feature intangible capital, a reduction in physical investment rate can only be attributed to a decline in depreciation rate.

4.7 Model misspecification

Finally, we explore model misspecification by simulating alternative models and reestimating the baseline model using moments generated from these alternatives. The baseline estimate will be robust to misspecification if it is close to these alternative estimates. This approach is similar to Catherine et al. (2022).

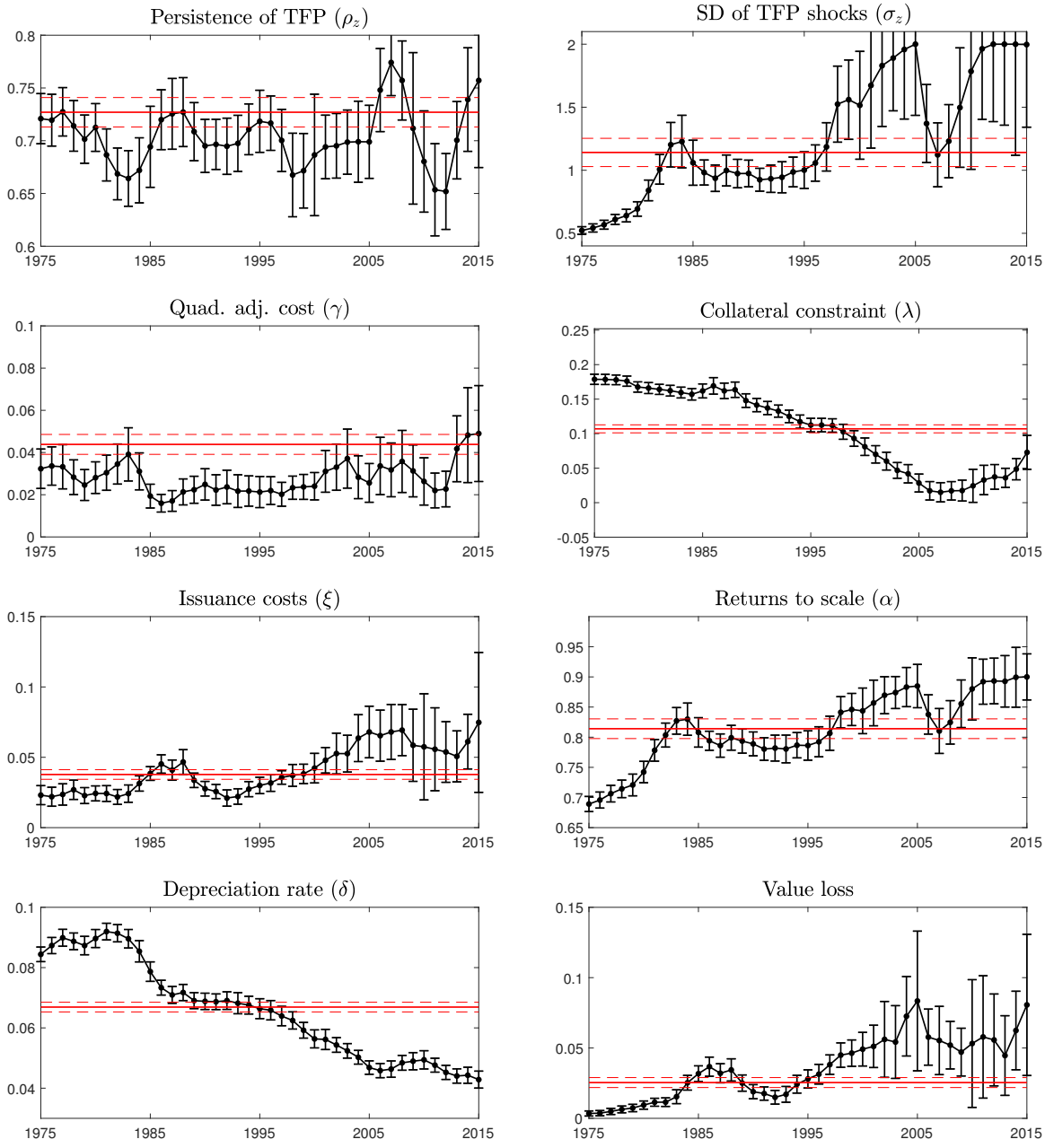
We illustrate our approach in the context of the recent corporate finance literature,

Figure 3: Histogram of parameter estimates across all identified sets of targeted moments (Corporate Finance Model)



Notes. This figure explores the sensitivity of parameter estimates to moment selection. Our baseline estimation targets seven moments $(m_i)_{i\in\{1..7\}}$. We consider a set of 10 additional moments used in the literature to estimate similar models: $(m_i)_{i\in\{8..17\}}$ described in Section 4.1 and Appendix B.1. We construct all possible sets of moments that contain the seven baseline moments and any combination of the 10 additional moments. These sets of moments are then used as targeted moments in an approximate SMM. This results in 1,024 sets of parameter estimates. After dropping cases where estimates are poorly-identified – where the standard errors for an estimated parameter is more than 10 times larger than the standard errors of the baseline estimate – we end up with 722 estimates. Each panel in the figure shows the distribution of parameters across these 722 estimations. The vertical black line and dashed lines show the baseline parameter estimates, together with their 95% confidence interval.

Figure 4: Time-series estimates (Corporate Finance Model)



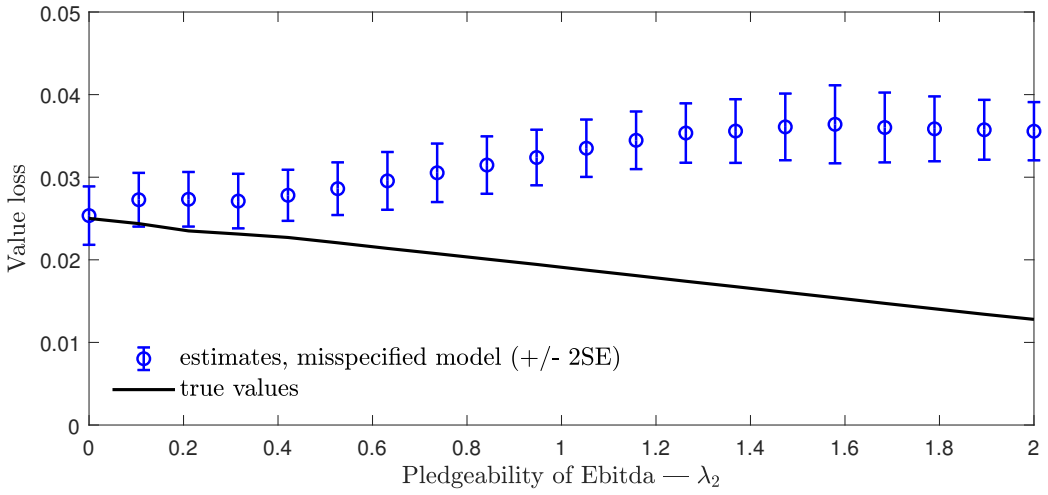
Notes. This figure shows the sensitivity of parameter estimates to the sample period used to compute the targeted moment in the data. For each year t on the x-axis, we compute the seven baseline moments $(m_i)_{i\in\{1..7\}}$ on the sample period $[t-5, t+4]$. We then reestimate the model using the benchmark approximate SMM that targets the moments measured for this sub-period. The y-axis reports the resulting parameter estimates and their 95% confidence interval. The horizontal solid and dashed red lines corresponds to the baseline estimates obtained when computing moments on the entire sample, together with their 95% confidence interval.

which shows that financial constraints often take the form of cash-flow constraints rather than collateral constraints (e.g. Lian and Ma, 2021; Greenwald, 2019). Our objective is to assess the robustness of the baseline estimates to this particular source of misspecification.

To do so, we augment the baseline model of Section 4.1 to introduce this new channel. We rewrite the debt constraint as: $d_{t+1} < \lambda k_{t+1} + \lambda_2 \cdot \mathrm{E}[e^{z_{t+1}(1-\alpha)}]k_{t+1}^{\alpha}$. We then simulate 40 versions of this alternative model, where baseline parameters are the baseline estimates of Table 1, and λ_2 increases uniformly from 0 (no misspecification) to 2 (large misspecification). This provides us with 40 sets of alternative moments.

We then estimate the baseline model (i.e., the model with $\lambda_2 = 0$) 40 separate times by targeting these 40 different sets of moments. Figure 5 plots the estimated value loss from financing constraints (y-axis) against the value of λ_2 in simulating the targeted moments (x-axis). The black line corresponds to the actual value loss in the correctly-specified model with both collateral and cash-flow constraints, which intuitively decreases as λ_2 increases.

Figure 5: Model mispecification: Estimating model with collateral constraints on data generated by cash-flow constraints (Corporate Finance Model)



Notes. This figure explores the sensitivity of estimation to model misspecification. We consider an augmented version of our corporate finance model where firms can pledge a multiple λ_2 of their EBITDA so that their debt constraint takes the form $d_{t+1} < \lambda k_{t+1} + \lambda_2 \cdot \mathrm{E}[e^{z_{t+1}(1-\alpha)}]k_{t+1}^{\alpha}$. We assume that the correctly specified model is the augmented model where the baseline parameters are set to their estimated value in Table 1 and λ_2 takes various values from 0 (the baseline model) to 2. For each possible value of λ_2, we reestimate the baseline parameter values using the misspecified model that targets these moments (simulated with the correctly-specified model). The x-axis plots the values of λ_2 used in each estimation. The blue circles report the value loss from financial constraint estimated with the approximation to the misspecified model while we target moments generated by the correctly-specified model. The black line reports the true value loss from financial constraint in the correctly-specified model.

Figure 5 shows that estimates of value loss (from financial constraints) are robust to values of λ_2 below 0.6. However, as λ_2 increases, the misspecified model wrongly infers increasing value losses (while the true value losses are in fact going down). A simple explanation is that as λ_2 increases, firms can avoid issuing equity (they issue debt) and the average equity to asset ratio in the economy goes down. The baseline model infers from this reduced equity issuance that the costs of equity issuance is going up, which leads to a large value loss from financial constraint. This analysis illustrates how we can leverage our methodology to evaluate the effect of specific model misspecification.

5 Application to a dynamic portfolio choice model

In this section we apply our method to a life-cycle portfolio choice model. The model is a variant of Catherine (2021), excluding housing investment. Households choose between investing in a risk-free asset and the stock market over their life cycle. Compared to Viceira (2001) and Cocco et al. (2005), the model includes countercyclical income risk and a realistic Social Security system.

5.1 Model, parameters and moments

Macroeconomic environment The riskfree log return is r and the stock market log return in year t is $s_t = s_{1,t} + s_{2,t}$. s_1 will be correlated with labor income, and follows a normal mixture distribution:

$$s_{1,t} = \begin{cases} s^-_{1,t} \sim \mathcal{N}(\mu^-_s, \sigma^2_{s_1}) & \text{with probability} \quad p_s \\ s^+_{1,t} \sim \mathcal{N}(\mu^+_s, \sigma^2_{s_1}) & \text{with probability} \quad 1 - p_s \end{cases} \tag{12}$$

where we impose $\mu^-_s < \mu^+_s$ and interpret μ^-_s as the expected log return in stock market crash years, and p_s their frequency. $s_{2,t}$ is normally distributed with variance $\sigma^2_{s_2}$.

The growth of the log national wage index l_1 is:

$$l_{1,t} - l_{1,t-1} = \mu_l + \lambda_{ls} s_{1,t} + \varepsilon_{l,t}, \tag{13}$$

where $\varepsilon_{l,t}$ follows $\mathcal{N}(0, \sigma^2_l)$, μ_l is the average growth rate, and λ_{ls} captures the correlation with stock returns.

Income risk Labor earnings L_{it} is the product of the national wage index $L_{1,t}$ and an idiosyncratic component $L_{2,it}$. The idiosyncratic component is further decomposed into a deterministic function of age φ_{it},[6] a persistent component z_{it} and a transitory shock η_{it}:

$$L_{2,it} = e^{\varphi_{it} + z_{it} + \eta_{it}}. \tag{14}$$

The persistent component follows an AR(1) process, with innovations drawn from a normal mixture. Specifically, the dynamics of z_i are given by

$$z_{it} = \rho_z z_{it-1} + \zeta_{it}, \tag{15}$$

where

$$\zeta_{it} = \begin{cases} \zeta_{it}^{-} \sim \mathcal{N}\left(\mu_{z,t}^{-}, \sigma_z^{-2}\right) & \text{with probability} \quad p_z \\ \zeta_{it}^{+} \sim \mathcal{N}\left(\mu_{z,t}^{+}, \sigma_z^{+2}\right) & \text{with probability} \quad 1 - p_z. \end{cases} \tag{16}$$

Here we impose $p_z \mu_{z,t}^{-} + (1 - p_z)\mu_{z,t}^{+} = 0$ and $p_z \leq 0.5$. The values of p_z, $\mu_{z,t}^{-}$ and $\mu_{z,t}^{+}$ control the degree of asymmetry in the distribution of income shocks. If $\sigma_z^{-} \gg \sigma_z^{+}$, p_z represents the frequency of significant events in a worker's career. To capture the cyclicality of skewness, $\mu_{z,t}^{-}$ is an affine function of the log growth rate of the wage index:

$$\mu_{z,t}^{-} = \overline{\mu_z^{-}} + \lambda_{zl}(l_{1,t} - l_{1,t-1}). \tag{17}$$

Finally, the transitory component of income is also modeled as a mixture of normals whose first and second components always coincide with the first and second components of the normal mixture governing the innovations to z_i.

$$\eta_{it} = \begin{cases} \eta_{it}^{-} \sim \mathcal{N}(0, \sigma_\eta^{-2}) & \text{if} \quad \zeta_{it} = \zeta_{it}^{-} \\ \eta_{it}^{+} \sim \mathcal{N}(0, \sigma_\eta^{+2}) & \text{if} \quad \zeta_{it} = \zeta_{it}^{+}. \end{cases} \tag{18}$$

Social Security Social Security payroll taxes represent 12.4% of the agent's earnings below the maximum taxable earnings, which represents 2.5 times the national wage index.

$$\mathrm{T}_{it} = .124 \cdot \min\left\{L_{it}, 2.5 \cdot L_{1,t}\right\}. \tag{19}$$

Retirement benefits depend on historical taxable earnings, adjusted for the growth

[6] Specically, we assume φ to be a polynomial function of age: $\varphi_2 \text{age}^2/100 + \varphi_1 \text{age}/10 + \varphi_0$.

in the national wage index. Specifically, the agent's Social Security benefits B are:

$$\frac{B_i}{L_{1,R}} = \begin{cases} .9 \cdot S_{iR} & \text{if } S_{iR} < .2 \\ .116 + .32 \cdot S_{iR} & \text{if } .2 \leq S_{iR} < 1 \\ .286 + .15 \cdot S_{iR} & \text{if } 1 \leq S_{iR}, \end{cases} \tag{20}$$

where R is the retirement age and $L_{1,R}$ is the value of the wage index at that age. The variable S_{it} keeps track of a worker's average taxable idiosyncratic earnings: $S_{it} = \sum_{k=t_0}^{t} \frac{\min\left\{L_{2,ik}, 2.5\right\}}{t-t_0+1}$, where t_0 denotes his first year of earnings.

Agent Given processes for labor, capital and social security incomes, the agent chooses consumption and the share of his wealth invested in the stock market portfolio to maximize the discounted sum of utilities from consumption C_t:

$$V_{t_0} = \mathrm{E} \sum_{t=t_0}^{T} \beta^{t-1} \left(\prod_{k=0}^{t-1} (1 - m_k) \right) \frac{C_t^{1-\gamma}}{1-\gamma}, \tag{21}$$

where γ is the coefficient of relative risk-aversion, m_k the mortality rate at age k, β the subjective discount factor and T the maximum lifespan. Mortality rates and age of death are calibrated. Finally, wealth evolves as:

$$W_{it+1} = \left[W_{it} + L_{it} + B_{it} - \mathrm{T}_{it} - C_{it} - c_{it} \right] \cdot \left[\pi_{it} e^{s_t} + (1 - \pi_{it}) e^{r} \right], \tag{22}$$

where π_{it} is the share of his wealth invested in equity. Owning stocks incurs a cost $c_{it} = \Phi L_{1,t}$ if $\pi_{it} > 0$. Short selling or leveraging are not allowed, such that $0 \leq \pi_{it} \leq 1$.

Moments and estimated parameters With other parameters calibrated (see Appendix Table C.1), we estimate three model parameters: the discount factor β, risk aversion γ, and the cost of owning equity Φ; thus, $\theta = (\gamma, \beta, \Phi)$. As in Catherine (2021), parameter estimation is based on matching three baseline moments from the Survey of Consumer Finances (1989–2016): (1) the wealth-to-labor income ratio, (2) the percentage of households holding equity, and (3) the stock share of wealth among equity holders. The construction of these moments is explained in Appendix C.2, and their role in identifying θ is discussed in Catherine (2021).

5.2 Training the aproximate moment function f_n

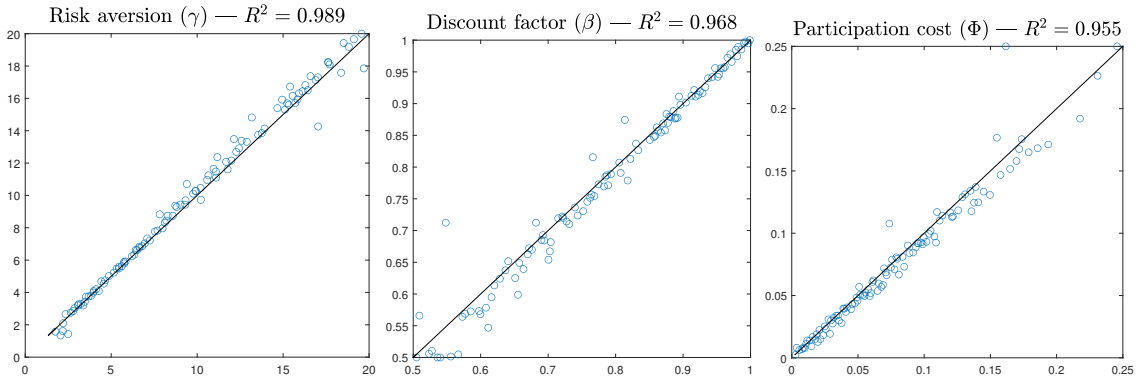
Similar to the corporate finance model, we build a training sample $\mathcal{D}_n$ of size 2,000 to train the approximate moment function f_n. The sample contains random parameter draws θ_i and the corresponding moments $f(\theta_i)$ obtained by simulating the model. We also construct an (out-of-sample) evaluation set of size 200, out of which 106 are well-identified parameters to assess the precision of the estimates by approximation f_n. Exact details about the construction of these samples are provided in Appendix C.3. Note that the training and evaluation sets are smaller than in our corporate finance example, so we can assess the effect of sample size on our methodology.

We train the approximate moment function f_n using the training sample and a neural network, a 5-layer MLP with 256, 128, 64, 32 and 16 nodes. Next section presents out-of-sample performance results.

5.3 Approximate moment estimation: Performance

The performance evaluation results are similar to the corporate finance model, as shown in Figure 6, the counterpart to Figure 1. For each draw in the evaluation set, the scatter plot compares the true parameters (x-axis) to the parameters estimated by targeting moments in the set using the approximate moment function f_n (y-axis). The out-of-sample R^2 is above 95% for all three parameters.

Figure 6: Out-of-sample performance (Household Finance Model)



Notes. This figure shows, for the household finance model, the precision in the out-of-sample evaluation set of our benchmark approximate SMM across estimated parameters. For each draw $(\theta, f(\theta))$ in the evaluation set, we use neural nets that construct the approximate moment function f_n and estimate parameters $\widehat{\theta}_n$ which match $f(\theta)$. The x-axis reports the true parameters θ, while the y-axis reports the estimated parameters $\widehat{\theta}_n$.

As in the corporate finance application, we also test our method on real data,

using the three baseline empirical moments from the SCF (see Appendix Table C.2, column 1). We estimate parameters using (1) actual SMM, (2) the approximate neural net moment function, and (3) the approximate moment corrected with the error estimate formula from (3). Table 2 shows that the approximate estimates are close to the true SMM. Both approaches give similar values for risk-aversion (γ about 8.3) and discount factor (β about .91). The participation cost Φ is estimated at .0053 in the true SMM, and .0056 in the approximate SMM, a small economic difference of about $50 a year. The third line in Table 2 confirms that the approximation error formula (3) usefully corrects the approximate SMM. Appendix Table C.2 reports close moment fits for the approximate SMM.

Table 2: Parameter estimates, true vs approximate SMM (Household Finance Model)

	γ	β	Φ
true SMM	8.328	0.9106	0.0053
- s.e., local deriv.	0.064	0.0021	0.0002
approx. SMM	8.621	0.9048	0.0056
- s.e., local fit deriv.	0.083	0.0024	0.0003
approx. SMM, corrected	8.371	0.9098	0.0052
- s.e., local fit deriv.	0.077	0.0023	0.0003
estimation, lower bound	1.01	0.5	0
estimation, upper bound	20	1	0.25

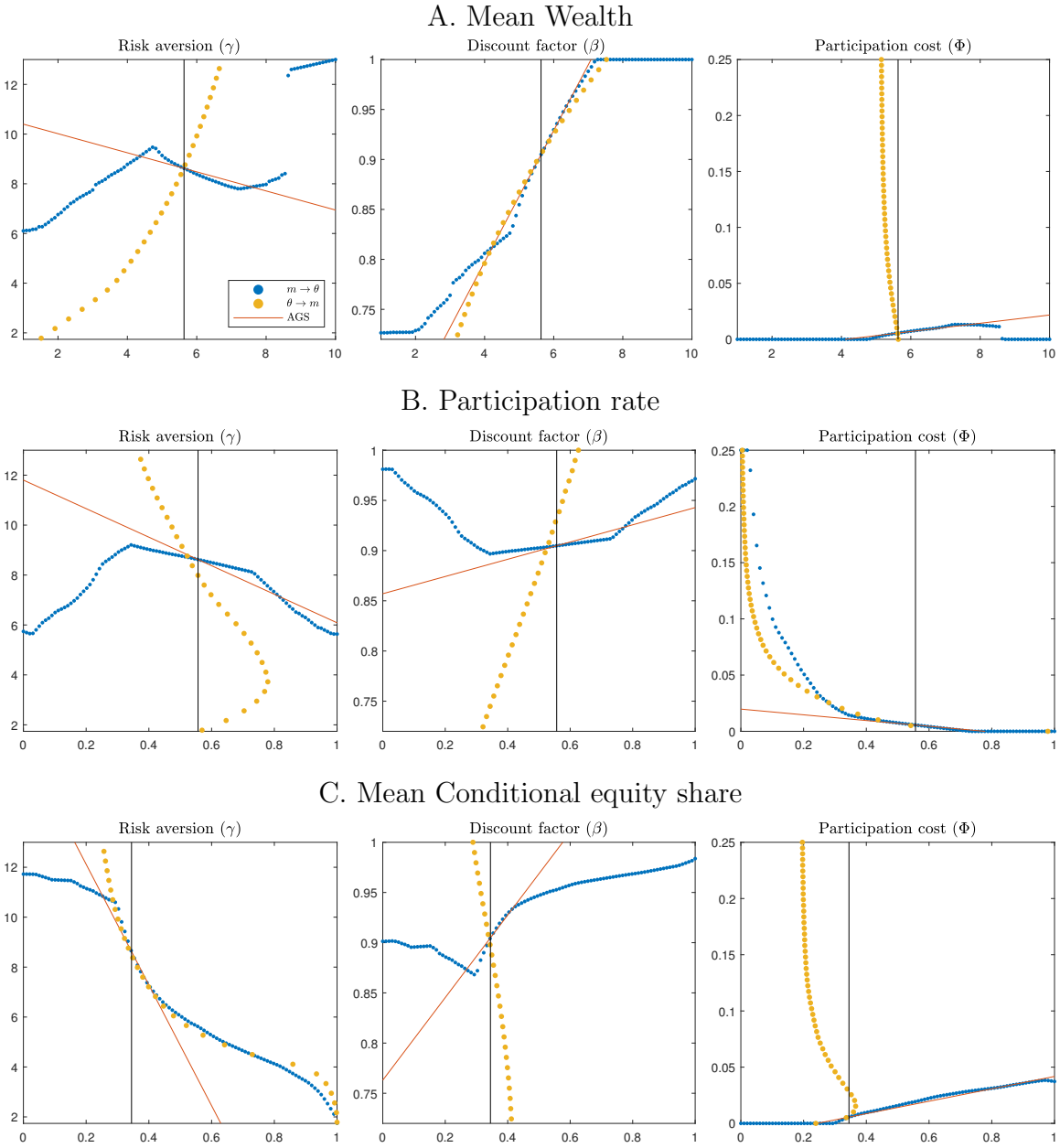
Notes. The table reports the parameter estimates of the household finance model presented in Section 5.1. We report parameter estimates using the true SMM (first line), the approximate SMM (second line), and the approximate SMM corrected with the first-order error formula from Section 3.3 (third line). The approximate SMM uses nerual net fit to approximate moment functions. γ is risk-aversion. β is the subjective discount factor. Φ is the participation cost.

5.4 Identification diagnostic

Figure 7, analog to Figure 2, shows how changes in targeted moments affect estimated parameters (blue line). This analysis is made possible by the low computational cost of the approximate SMM. The figure includes nine panels, each representing a moment-parameter pair.

Similar takeaways to the corporate finance model emerge from this analysis. First, relying on local comparative statics from $f(\theta)$ can be misleading. For instance, in the top left panel, comparative statics suggest that higher mean wealth results from

Figure 7: Sensitivity of parameters to moments (Household Finance Model)



Notes. This figure plots parameter values on the y-axis and the value of moment on the x-axis. Panel A (resp. B and C) corresponds to mean wealth (resp. participation rate and conditional equity share). The yellow line draws the function $f(\theta)$ – moments as functions of parameter values. The blue line plots how variations in moment values – holding other moments fixed – affect parameter estimates. Each dot on the blue line corresponds to a separate estimation using our approximation technique. The red line corresponds to the local linear approximation of the blue line around the parameter estimates, i.e. the "sensitivity matrix" of Andrews et al. (2017). The black vertical line indicates the value of a moment in the data.

increased risk aversion (yellow line), as more risk-averse households save more. However, the actual estimation (blue line) shows the opposite: holding the participation rate and equity share constant, the model needs to raise discount factor to match the higher mean wealth, and *reduce* risk aversion to offset the induced increase in household savings with equity investment.

Second, the linear approximation from Andrews et al. (2017) can fail when considering moment values that differ significantly from their empirical values. For example, an equity share of 0.6 (instead of 0.35 in the SCF) would lead to a risk-aversion of 5 in our estimation; the linear approximation would estimate it at 2 instead.

5.5 Robustness to moment selection

As in our corporate finance application, we analyze the robustness of parameter estimates to moment selection. To do this, we expand our set of moments from the initial three to include an additional 64 moments. The first group of four moments deals with normalization and non-normality: (1) median wealth, (2) median conditional equity share, (3) an alternative definition of mean conditional equity share, normalized by financial wealth rather than net worth, and (4) the median of this alternative conditional equity share. The first two moments address the fat upper tails of stock holdings and total wealth, while the last two adjust for differences between net worth and financial wealth, which are identical in the model but not in the data.

The second group consists of 60 moments, which are the baseline moments (mean wealth, stock participation, and equity share) broken down by 20 age groups, ranging from ages 23–25 to 80–82.[7] The procedure is explained in detail in Catherine (2021).

We then reestimate the model using 72 alternative *sets* of moments. Unlike the corporate finance case, where all combinations were explored, the portfolio choice model cannot simultaneously match equity shares normalized by both total wealth and financial wealth. Thus, we focus on the following moment combinations:

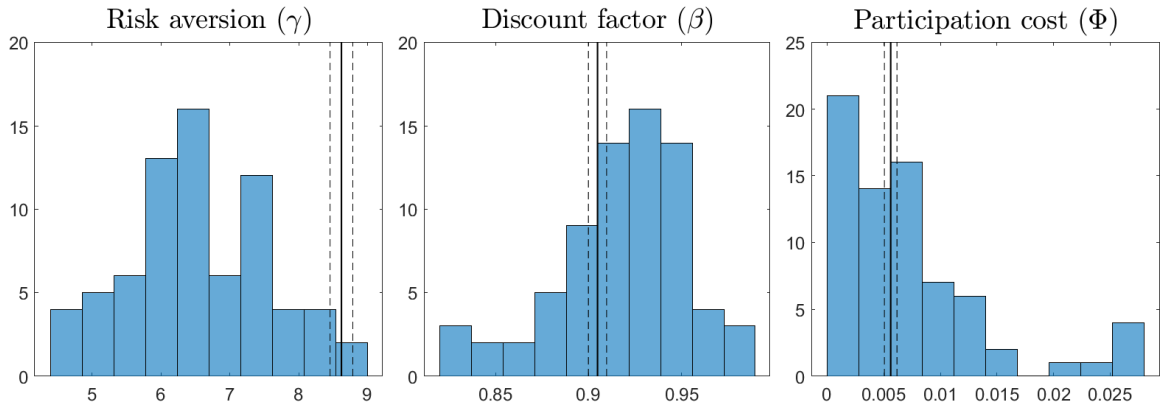
- 12 combinations of the three baseline moments: (a) mean wealth (overall or by age group), (b) mean participation (overall or by age group), and (c) mean equity share (overall, by age group, or normalized by financial wealth).

[7] We compute the life-cycle of *unconditional* equity share, instead of conditional equity share, as the model often returns zero participation for some age groups, making the conditional mean equity share undefined.

- 36 combinations that add *either* (a) median wealth, (b) median equity share, or (c) median equity share normalized by financial wealth to the previous 12.
- 24 combinations that add *both* median wealth and one of the two median equity share measures to the previous combinations.

Figure 8 shows the distribution of parameter estimates across all 72 sets of moments. Most combinations produce consistent risk-aversion estimates, with the distribution peaking at 6–7, although this peak is substantially lower than the baseline estimate of 8.5. The discount factor (β) estimates range from 0.9 to 0.96, closely clustering around the baseline estimate of 0.93. For participation costs, the distribution peaks at 0, significantly lower than the baseline estimate of 0.005.

Figure 8: Histogram of parameter estimates across all selected sets of targeted moments (Household Finance Model)



Notes. This figure explores the sensitivity of parameter estimates to moment selection. Our baseline estimation targets three moments $(m_i)_{i\in\{1..3\}}$. We consider 72 alternative set of moments described in Section 5.5. Each panel in the figure shows the distribution of parameters across these 72 estimations. The vertical black line and dashed lines report the baseline parameter estimates, together with their 95% confidence interval.

6 Conclusion

This paper presents a fast and straightforward method for conducting robustness checks and identification diagnostics in structural estimation. Our approach involves estimating an approximation of moments as a fucntion of model parameters using a training dataset. Once this approximation is fitted, it allows for the rapid estimation of structural parameters. In our two applications—corporate finance and household

finance models—we demonstrate that this "approximate SMM" drastically reduces computational costs compared to standard SMM, with minimal loss of precision.

This reduction in computational burden opens up three important exercises that were previously challenging to conduct. First, we can now easily assess parameter robustness to moment selection. Second, sample-split analyses become feasible, allowing for quicker assessments of model robustness and validity across different subsamples. Lastly, the approximate SMM enables us to explore the sensitivity of baseline estimates to misspecification bias. By simulating various alternative models, we can evaluate how deviations from the baseline model affect parameter estimates.

References

Abel, A. B. and J. C. Eberly (1994, December). A unified model of investment under uncertainty. *American Economic Review 84*(5), 1369–1384.

Ameriks, J., J. Briggs, A. Caplin, M. D. Shapiro, and C. Tonetti (2020). Long-Term-Care Utility and Late-in-Life Saving. *Journal of Political Economy 128*(6), 2375–2451.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017). Measuring the sensitivity of parameter estimates to estimation moments. *Quarterly Journal of Economics 132*(4).

Andrews, I., M. Gentzkow, and J. M. Shapiro (2020a). On the informativeness of descriptive statistics for structural estimates. *Econometrica (Matthew Gentzkow's Fisher-Schultz Lecture) 88*(6), 2231–2258. Working Paper.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2020b). Transparency in structural research. *Journal of Business and Economic Statistics (invited discussion paper) 38*(4), 711–722.

Armstrong, T. B. and M. Kolesár (2021, January). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

Azinovic, M., L. Gaegauf, and S. Scheidegger (2019). Deep equilibrium nets.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning 14*, 115–133.

Bates, T., K. Kahle, and R. Stulz (2009). Why do u.s. firms hold so much more cash than they used to? *Journal of Finance 64*, 1985–2021.

Begenau, J. (2020). Capital requirements, risk choice, and liquidity provision in a business-cycle model. *Journal of Financial Economics 136*(2), 355–378.

Berger, D., K. Herkenhoff, and S. Mongey (2025). Minimum wages, efficiency, and welfare. *Econometrica 93*(1), 265–301.

Bonhomme, S. and M. Weidner (2018, July). Minimizing Sensitivity to Model Misspecification. Papers 1807.02161, arXiv.org.

Catherine, S. (2021, 12). Countercyclical labor income risk and portfolio choices over the life cycle. *The Review of Financial Studies 35*(9), 4016–4054.

Catherine, S., T. Chaney, Z. Huang, D. Sraer, and D. Thesmar (2022). Quantifying reduced-form evidence on collateral constraints. *Journal of Finance*.

Chen, H., A. Didisheim, and S. Scheidegger (2021, February). Deep Structural Estimation:With an Application to Option Pricing. Cahiers de Recherches Economiques du Département d'économie 21.14, Université de Lausanne, Faculté des HEC, Département d'économie.

Cocco, J. F., F. J. Gomes, and P. J. Maenhout (2005, 02). Consumption and Portfolio Choice over the Life Cycle. *The Review of Financial Studies 18*(2), 491–533.

Creel, M. (2021). Inference using simulated neural moments. *Econometrics 9*(4), 35.

Creel, M., J. Gao, H. Hong, and D. Kristensen (2015). Bayesian indirect inference and the abc of gmm. *arXiv preprint arXiv:1512.07385 2*.

Crouzet, N. and J. Eberly (2018). Intangibles, investment, and efficiency. *American Economic Review: AEA Papers and Proceedings 108*, 426–431.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems 2*(4), 303–314.

Duarte, V. (2020). Machine learning for continuous-time finance.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Fernandez-Villaverde, J., M. E. Kahou, J. Perla, and A. Sood (2021). Exploiting symmetry in high-dimensional dynamic programming.

Fernández-Villaverde, J., G. Nuño, G. Sorg-Langhans, and M. Vogler (2021). A deep learning algorithm for high-dimensional dynamic programming problems.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications* (2nd ed.). New York: Wiley.

Fonseca, J., V. Duarte, A. Goodman, and J. Parker (2022). Benchmarking Global Optimizers. NBER Working Papers 29559.

Gomes, F. (2020, December). Portfolio Choice Over the Life Cycle: A Survey. *Annual Review of Financial Economics 12*(1), 277–304.

Gouriéroux, C. and A. Montfort (1996). *Simulation-Based Econometric Methods*.

Greenwald, D. (2019). Firm debt covenants and the macroeconomy: The interest coverage channel. Technical report.

Greenwald, D. L. and A. Guren (2021, October). Do credit conditions move house prices? Working Paper 29391, National Bureau of Economic Research.

Guvenen, F. and A. A. Smith (2014, November). Inferring Labor Income Risk and Partial Insurance From Economic Choices. *Econometrica 82*, 2085–2129.

Hennessy, C. A. and T. M. Whited (2007). How costly is external financing? evidence from a structural estimation. *The Journal of Finance 62*(4), 1705–1745.

Huber, P. J. (2011). *Robust Statistics*, pp. 1248–1251. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jones, D. R. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization 21*(4), 345–383.

Lang, S. (2012). *Real and Functional Analysis*. Graduate Texts in Mathematics. Springer New York.

Li, Q. and J. S. Racine (2023). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Lian, C. and Y. Ma (2021). Anatomy of corporate borrowing constraints. *Quarterly Journal of Economics 136*.

Maliar, L., S. Maliar, and P. Winant (2019, September). Will Artificial Intelligence Replace Computational Economists Any Time Soon? CEPR Discussion Papers 14024, C.E.P.R. Discussion Papers.

Midrigan, V. and D. Y. Xu (2014). Finance and misallocation: Evidence from plant-level data. *American Economic Review 104*(2), 422–58.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Norets, A. (2012). Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations. *Econometric Reviews 31*(1), 84–106.

Piazzesi, M., M. Schneider, and J. Stroebel (2020, March). Segmented housing search. *American Economic Review 110*(3), 720–59.

Rudin, W. (1953). *Principles of mathematical analysis*.

Viceira, L. M. (2001). Optimal portfolio choice for long-horizon investors with non-tradable labor income. *The Journal of Finance 56*(2), 433–470.

Villa, A. T. and V. Valaitis (2019). Machine learning projection methods for macro-finance models. *International Political Economy: Investment & Finance eJournal*.

APPENDIX – FOR ONLINE PUBLICATION

A Proofs

This section is dedicated to proofs. The necessary conventions are outlined in Appendix A.1 and the main proofs are presented in the subsequent sections.

A.1 Conventions

For definitions and basic facts, we refer to Rudin (1953); Folland (1999); Lang (2012). Below, we provide the basic notation.

The transpose is denoted by $\cdot^\top$, and vectors $x \in \mathbb{R}^d$ are treated as $d \times 1$ columns. We write $\mathbb{N}$ for the positive integers and $\mathbb{R}$ for the reals. Set difference is $\mathbf{A} \setminus \mathbf{B}$, and Cartesian products are $A \times B$. For a set $\mathbf{A}$, $\mathrm{vol}(\mathbf{A})$ denotes its volume, and its diameter is $\mathrm{diam}(\mathbf{A}) = \sup_{x,y\in\mathbf{A}} \|x - y\|_2$. The closed and open balls around θ with radius r are denoted by $\bar{\mathbf{B}}(\theta, r)$ and $\mathbf{B}(\theta, r)$, respectively. Finally, uniform convergence of functions is written as $h_i \xrightarrow{\text{unif}} h$.

A.2 Proof of Theorem 1

Proof. Consider the following notation of (2):

$$\Theta_n = \arg\min_{\theta\in\mathbf{P}_\theta} h_n(\theta), \quad \text{and} \quad h_n(\theta) = (m - f_n(\theta))^\top W (m - f_n(\theta)), \tag{A.1}$$

Here Θ_n denotes the set of all solutions to the minimization. Let $h(\theta) = (m - f(\theta))^\top W (m - f(\theta))$. The goal is to show that, for any choice of $\hat{\theta}_n \in \Theta_n$, we have $\hat{\theta}_n \to \theta^*$. To this end, we adapt a non-random version of Newey and McFadden (1994, Theorem 2.1) to our setting. We first show that $h_n \xrightarrow{\text{unif}} h$ on $\mathbf{P}_\theta$. To see this, observe that h_n is formed through basic operations on the components of f_n, and these components are uniformly convergent to those of f and are bounded.

As the next step, given $\epsilon > 0$, we show that there exists N such that $h(\hat{\theta}_n) < h(\theta^*) + \epsilon$ for all $n > N$. Since $\hat{\theta}_n$ minimizes h_n on $\mathbf{P}_\theta$, we have $h_n(\hat{\theta}_n) < h_n(\theta^*) + \epsilon/3$. Moreover, since $h_n \xrightarrow{\text{unif}} h$, there exists N such that for all $n > N$, both $h(\hat{\theta}_n) < h_n(\hat{\theta}_n) + \epsilon/3$ and $h_n(\theta^*) < h(\theta^*) + \epsilon/3$ hold. Combining these inequalities yields $h(\hat{\theta}_n) < h(\theta^*) + \epsilon$ for all $n > N$.

Finally, to prove $\theta_n \to \theta^*$, given an open neighbourhood $\mathbf{O}$ of θ^*, it suffices to show that $\hat{\theta}_n$ lies in $\mathbf{O}$ for all sufficiently large n. To this end, consider the set $\mathbf{P}_\theta \setminus \mathbf{O}$, which is compact. Since h is continuous, $\inf_{\theta\in\mathbf{P}_\theta\setminus\mathbf{O}} h(\theta) = h(\bar{\theta})$ exists for some $\bar{\theta} \in \mathbf{P}_\theta$. Because θ^* is the unique minimizer of h, we have $h(\theta^*) < h(\bar{\theta})$. Choosing $\epsilon = h(\bar{\theta}) - h(\theta^*)$ in Step 2 then yields $h(\hat{\theta}_n) < h(\bar{\theta})$ for $n > N$. This implies $h(\hat{\theta}_n) < \inf_{\theta\in\mathbf{P}_\theta\setminus\mathbf{O}} h(\theta)$, so $\hat{\theta}_n$ cannot lie in $\mathbf{P}_\theta \setminus \mathbf{O}$. Hence, $\hat{\theta}_n \in \mathbf{O}$. □

A.3 Proof of Proposition 1

Proof. Since $\mathbf{O}$ is open, there exists an open ball $\mathbf{B}$ centered at θ^* such that $\mathbf{B} \subseteq \mathbf{O}$. Next, Theorem 1 implies that $\hat{\theta}_n \to \theta^*$. Therefore, there exists N_1 such that $\hat{\theta}_n \in \mathbf{B}$ for $n \geq N_1$. For the rest of the proof, suppose $n \geq \max(N_0, N_1)$. Since ∇f_n exists and is continuous on $\mathbf{B} \subseteq \mathbf{O}$, the derivative of the objective function of the minimization problem in (2) exists and is continuous on $\mathbf{B}$, and thus it vanishes at the corresponding minimizer $\hat{\theta}_n \in \mathbf{B}$. Therefore,

$$\nabla f_n(\hat{\theta}_n)^\top W \left(f(\theta^*) - f_n(\hat{\theta}_n)\right) = 0. \tag{A.2}$$

Moreover, since $\hat{\theta}_n \to \theta^*$ and ∇f exists and is continuous on an open and convex set $\mathbf{B} \subseteq \mathbf{P}_\theta$ containing $\hat{\theta}_n$, we can write the following first-order Taylor expansion of f in the vicinity of $\hat{\theta}_n$:

$$f(\theta^*) = f(\hat{\theta}_n) - \nabla f(\hat{\theta}_n)(\hat{\theta}_n - \theta^*) + \mathcal{R}, \tag{A.3}$$

where $\mathcal{R} = \mathcal{O}(\|\hat{\theta}_n - \theta^*\|_2^2)$ because the second derivative of f is bounded. Next, replacing $f(\theta^*)$ from (A.3) into (A.2) results in:

$$\nabla f_n(\hat{\theta}_n)^\top W \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n) - \nabla f(\hat{\theta}_n)(\hat{\theta}_n - \theta^*) + \mathcal{R} \right) = 0. \tag{A.4}$$

Re-arranging the terms in (A.4) leads to:

$$\left(\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n) \right) \left(\hat{\theta}_n - \theta^* \right) = \nabla f_n(\hat{\theta}_n)^\top W \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n) + \mathcal{R} \right). \tag{A.5}$$

The fact that Λ_n exists means that $\nabla f_n(\hat{\theta}_n)^\top W \nabla f(\hat{\theta}_n)$ is invertible. This allows us to simplify (A.5) into the following:

$$\hat{\theta}_n - \theta^* = \Lambda_n \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n) \right) + \Lambda_n \mathcal{R}. \tag{A.6}$$

Finally, having $\|\Lambda_n\|_2 \leq L_0$ for all $n \geq N_0$, we conclude that $\Lambda_n \mathcal{R} = \mathcal{O}(\|\hat{\theta}_n - \theta^*\|_2^2)$. □

A.4 Proof of Lemma 1

We first prove an intermediate lemma, upon which the main proof is subsequently built.

Lemma A.1. *Let $\mathbf{P} = [a_1, b_1] \times \ldots \times [a_K, b_K] \subset \mathbb{R}^K$, with $a_i < b_i < \infty$ for $1 \leq i \leq K$. Then, there exists a constant $\xi(\mathbf{P}) > 0$ such that for all $x \in \mathbf{P}$ and all $0 \leq r \leq \operatorname{diam}(\mathbf{P})$, we have:*

$$\operatorname{vol}\left(\mathbf{B}(x, r) \cap \mathbf{P}\right) \geq \xi(\mathbf{P}) \cdot \operatorname{vol}\left(\mathbf{B}(x, r)\right). \tag{A.7}$$

Proof of Lemma A.1. For sufficiently small r_0, if $r < r_0$, the quantity $\operatorname{vol}(\mathbf{B}(x, r) \cap \mathbf{P})$ attains its minimum when x is a vertex of the box $\mathbf{P}$, and this minimum equals $2^{-K} \operatorname{vol}(\mathbf{B}(x, r))$. When $r \geq r_0$, we consider the following minimization problem:

$$\begin{aligned} \xi(\mathbf{P}) = \inf \quad & \frac{\operatorname{vol}\left(\mathbf{B}(x, r) \cap \mathbf{P}\right)}{\operatorname{vol}\left(\mathbf{B}(x, r)\right)}, \\ \text{s.t.} \quad & x \in \mathbf{P}, \\ & r \in [r_0, \operatorname{diam}(\mathbf{P})]. \end{aligned} \tag{A.8}$$

Note that (A.8) is a minimization of a continuous function over a compact set, so an optimal solution is attained at some $(\bar{x}, \bar{r}) \in \mathbf{F}$. Letting $\xi(\mathbf{P}) = f(\bar{x}, \bar{r})$ completes the proof. □

Proof of Lemma 1. Since the data points are drawn randomly, the quantity δ_n is a random variable, and the goal is to find a probabilistic bound on δ_n. To this end, we start by investigating the probability of $\delta_n > r$ for a given $r > 0$. Note that if $\delta_n > r$, then the definition of δ_n in (4) implies that there exists $\alpha_0 \in \mathbf{P}_\theta$ such that no data point lies in $\mathbf{B}(\alpha_0, r)$. As a result, we can write the following inequality for the probability of $\delta_n > r$:

$$\mathbb{P}\left[\delta_n > r\right] \leq \mathbb{P}\left[\text{no data point lies in } \mathbf{B}(\alpha_0, r)\right] = \left(1 - \frac{\operatorname{vol}\left(\mathbf{B}(\alpha_0, r) \cap \mathbf{P}_\theta\right)}{\operatorname{vol}(\mathbf{P}_\theta)}\right)^n. \tag{A.9}$$

Therefore, due to Lemma A.1, there exists $\xi(\mathbf{P}_\theta) > 0$ such that:

$$\operatorname{vol}\left(\mathbf{B}(\alpha_0, r) \cap \mathbf{P}_\theta\right) \geq \xi(\mathbf{P}_\theta) \cdot \operatorname{vol}\left(\mathbf{B}(\alpha_0, r)\right). \tag{A.10}$$

Next, let $p = \mathbb{P}[\delta_n > r]$. By substituting this together with (A.10) into (A.9), we conclude that:

$$p \leq \left(1 - \frac{\xi(\mathbf{P}_\theta) \cdot \operatorname{vol}(\mathbf{B}(\alpha_0, r))}{\operatorname{vol}(\mathbf{P}_\theta)}\right)^n \quad \Rightarrow \quad \log p \leq n \log\left(1 - c_0\, r^K\right), \tag{A.11}$$

where $c_0 > 0$ is a constant that does not depend on r, n, or p. Next, using the Taylor series expansion of

$\log(1/(1-x))$ for $x \in (0,1)$, the following inequality results from (A.11):

$$\log \frac{1}{p} \geq n \log \frac{1}{1 - c_0\, r^K} = n \left(c_0\, r^K + \sum_{i=2}^{\infty} \frac{(c_0\, r^K)^i}{i} \right) \geq n c_0 r^K. \tag{A.12}$$

Therefore,

$$r \leq c_0^{-\frac{1}{K}} \cdot n^{-\frac{1}{K}} \cdot \log^{\frac{1}{K}} \left(\frac{1}{p} \right). \tag{A.13}$$

Recalling $p = \mathbb{P}(\delta_n > r)$, we know that with probability $1-p$, we have $\delta_n \leq r$. This, together with (A.13), leads to the fact that with probability $1-p$, $\delta_n = (-\log p)^{1/K} \cdot \mathcal{O}(n^{-1/K})$. □

A.5 Proof of Proposition 2

Proof. Let f^r and f_n^r denote the r-th components, and let θ_i denote the i-th data point. Observe that $f_n(\theta)$ averages only the data points effective at θ, namely those lying in $\mathbf{B}(\theta, \lambda_n)$, since $k_n(\theta, \theta_i) = 0$ outside this ball. Therefore,

$$f_n(\theta) = \frac{\sum_{i \in \mathbf{I}_n(\theta)} k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i \in \mathbf{I}_n(\theta)} k_n(\theta, \theta_i)}, \tag{A.14}$$

where $\mathbf{I}_n(\theta) = \{1 \leq i \leq n \mid \theta_i \in \mathbf{B}(\theta, \lambda_n)\}$. Note that $\mathbf{I}_n(\theta) \neq \emptyset$ for $\theta \in \mathbf{P}_\theta$ because otherwise we would have $\min_{1 \leq i \leq n} |\theta - \theta_i|_2 \geq \lambda_n > \delta_n$. Now, (A.14) immediately implies:

$$\min_{i \in \mathbf{I}_n(\theta)} f^r(\theta_i) \leq f_n^r(\theta) \leq \max_{i \in \mathbf{I}_n(\theta)} f^r(\theta_i). \tag{A.15}$$

Next, assuming $n \in \mathbb{N}$, $1 \leq r \leq M$, and $\theta \in \mathbf{P}_\theta$ are given, we claim that there exist $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n)$ such that:

$$|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)|. \tag{A.16}$$

To prove (A.16), let:

$$\omega_1 = \operatorname*{arg\,min}_{\alpha \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta} f^r(\alpha), \qquad \omega_2 = \operatorname*{arg\,max}_{\alpha \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta} f^r(\alpha). \tag{A.17}$$

Note that ω_1 and ω_2 exist because f^r is continuous and $\bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$ is compact. Additionally, since $\theta \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$, we can write:

$$f^r(\omega_1) \leq f^r(\theta) \leq f^r(\omega_2). \tag{A.18}$$

As the next step, observe that if $i \in \mathbf{I}_n(\theta)$, then $\theta_i \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta$. Therefore,

$$f^r(\omega_1) = \min_{\alpha \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta} f^r(\alpha) \leq \min_{i \in \mathbf{I}_n(\theta)} f^r(\theta_i), \tag{A.19}$$

$$\max_{i \in \mathbf{I}_n(\theta)} f^r(\theta_i) \leq \max_{\alpha \in \bar{\mathbf{B}}(\theta, \lambda_n) \cap \mathbf{P}_\theta} f^r(\alpha) = f^r(\omega_2). \tag{A.20}$$

Putting (A.15), (A.19), and (A.20) together leads to:

$$f^r(\omega_1) \leq f_n^r(\theta) \leq f^r(\omega_2). \tag{A.21}$$

Finally, (A.18) and (A.21) yield (A.16). Having established (A.16), we proceed with the proofs as follows:

Part i. Using Theorem 1, it suffices to prove $f_n \xrightarrow{\text{unif}} f$. Given $\epsilon > 0$, note that each f^r is continuous on the compact set $\mathbf{P}_\theta$ and therefore uniformly continuous. Hence, there exists $\beta_r > 0$ such that $\|\alpha_1 - \alpha_2\|_2 < \beta_r$ implies $|f^r(\alpha_1) - f^r(\alpha_2)| < \epsilon/\sqrt{M}$. Let $\beta = \min(\beta_1, \ldots, \beta_M)$. Since $\lambda_n \to 0$, there exists N such that for all $n \geq N$, $\lambda_n < \beta/2$.

Next, by (A.16), there exist $\omega_1, \omega_2 \in \bar{\mathbf{B}}(\theta, \lambda_n) \subset \mathbf{B}(\theta, \beta/2)$ such that $|f_n^r(\theta) - f^r(\theta)| \leq |f^r(\omega_1) - f^r(\omega_2)|$.

Since $\|\omega_2-\omega_1\|_2<\beta$, we have $|f^r(\omega_2)-f^r(\omega_1)|<\epsilon/\sqrt{M}$. Hence, $|f_n^r(\theta)-f^r(\theta)|<\epsilon/\sqrt{M}$. Applying this to all components yields $\|f_n(\theta)-f(\theta)\|_2<\epsilon$ for all $\theta\in\mathbf{P}_\theta$ whenever $n\geq N$.

Part ii. We proceed through the following steps:

Step 1. Showing that ∇f_n exists and is continuous over any open set $\mathbf{O}\subseteq\mathbf{P}_\theta$:

Suppose the open set $\mathbf{O}\subseteq\mathbf{P}_\theta$ is given. Considering f_n from (6), it is straightforward to compute its corresponding $M\times K$ Jacobian matrix evaluated at $\theta\in\mathbf{O}$, which is given by:

$$\nabla f_n(\theta)=\frac{1}{\sum_{i=1}^n k_n(\theta,\theta_i)}\left[\sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}\, f(\theta_i)\,\nabla k_n(\theta,\theta_j)^\top\right], \tag{A.22}$$

where $f(\theta_i)\in\mathbb{R}^M$ is a column vector (i.e., it is $M\times 1$) and:

$$\alpha_{ij}=\begin{cases}-w_i & \text{if } i\neq j\\ 1-w_i & \text{if } i=j\end{cases},\quad w_i=\frac{k_n(\theta,\theta_i)}{\sum_{i=1}^n k_n(\theta,\theta_i)},\quad \nabla k_n(\theta,\theta_i)=\frac{2(\theta-\theta_i)}{\lambda_n^2}\eta'\left(\frac{\|\theta-\theta_i\|_2^2}{\lambda_n^2}\right). \tag{A.23}$$

Note that in (A.23), $\eta'(\cdot)$ denotes the derivative of $\eta(\cdot)$ and $\nabla k_n(\theta,\theta_i)$ is a $K\times 1$ vector. Moreover, (A.22) and (A.23) together imply that $\nabla f_n(\theta)$ exists and is continuous over $\mathbf{O}$, due to the existence and continuity of $\eta'(\cdot)$.

Step 2. Proving f is Lipschitz continuous around θ^*:

We know that the assumptions of Proposition 1 hold. As a result, there exists $\mathbf{O}\subseteq\mathbf{P}_\theta$ around θ^* such that ∇f is continuous over $\mathbf{O}$. In particular, ∇f is continuous at $\theta^*\in\mathbf{O}$. We assumed $\|\nabla f(\theta^*)\|_2\neq 0$, so there exists $\beta_0>0$ such that for all $\theta\in\mathbf{B}(\theta^*,\beta_0)\subseteq\mathbf{O}$, $\|\nabla f(\theta)-\nabla f(\theta^*)\|_2<\frac{1}{2}\|\nabla f(\theta^*)\|_2$, and consequently $\|\nabla f(\theta)\|_2<\frac{3}{2}\|\nabla f(\theta^*)\|_2$.

Let $L_1=\frac{3}{2}\|\nabla f(\theta^*)\|_2$ and define $\mathbf{S}=\mathbf{B}(\theta^*,\beta_0)$. Observe that $\mathbf{S}$ is open and convex in $\mathbb{R}^K$, and $\|\nabla f(\theta)\|_2<L_1$ holds for $\theta\in\mathbf{S}$. This leads to f being Lipschitz continuous with the Lipschitz factor L_1, due to the Mean Value Inequality.

Step 3. Finding the convergence rate of $\|\hat{\theta}-\theta^*\|_2$:

Recall $\mathbf{S}=\mathbf{B}(\theta^*,\beta_0)$ from Step 2. We know that $\lambda_n=\gamma\,\delta_n\to 0$. Therefore, there exists N_1 such that for $n\geq N_1$, we have $\lambda_n<\beta_0/2$. Now, suppose $n\geq N_1$, $\theta\in\mathbf{B}(\theta^*,\beta_0/2)$, and $1\leq r\leq M$ are given. Then $\bar{\mathbf{B}}(\theta,\lambda_n)\subseteq\mathbf{S}$. Additionally, from (A.16), there exist $\omega_1,\omega_2\in\bar{\mathbf{B}}(\theta,\lambda_n)$ such that $|f_n^r(\theta)-f^r(\theta)|\leq|f^r(\omega_1)-f^r(\omega_2)|$. This, together with the Lipschitz continuity of f on $\mathbf{S}$ from Step 2 and the fact that $\omega_1,\omega_2\in\mathbf{S}$, leads to the following inequality:

$$|f_n^r(\theta)-f^r(\theta)|\leq|f^r(\omega_1)-f^r(\omega_2)|\leq\|f(\omega_1)-f(\omega_2)\|_2\leq L_1\|\omega_1-\omega_2\|_2\leq 2L_1\lambda_n. \tag{A.24}$$

Note that the rightmost inequality in (A.24) holds because $\omega_1,\omega_2\in\bar{\mathbf{B}}(\theta,\lambda_n)$. By applying (A.24) to every component for $1\leq r\leq M$, we conclude that $\|f_n(\theta)-f(\theta)\|_2\leq 2\sqrt{M}L_1\lambda_n$ holds for $n\geq N_1$ and $\theta\in\mathbf{B}(\theta^*,\beta_0/2)\subseteq\mathbf{S}\subseteq\mathbf{O}$.

Next, having $\hat{\theta}_n\to\theta^*$ from part (i) of the theorem, there exists N_2 such that if $n\geq N_2$, then $\hat{\theta}_n\in\mathbf{B}(\theta^*,\beta_0/2)$. This implies that $\|f_n(\theta)-f(\theta)\|_2\leq 2\sqrt{M}L_1\lambda_n$ holds when $\theta=\hat{\theta}_n$. Therefore, knowing that the conditions of Proposition 1 are all met (due to Step 1 and the assumptions) and using its notation, we can write the following inequality for $n\geq\max(N_0,N_1,N_2)$:

$$\left\|\Lambda_n\left(f(\hat{\theta}_n)-f_n(\hat{\theta}_n)\right)\right\|_2\leq\|\Lambda_n\|_2\|f(\hat{\theta}_n)-f_n(\hat{\theta}_n)\|_2\leq L_0\cdot 2\sqrt{M}L_1\lambda_n=2L_0L_1\gamma\sqrt{M}\delta_n. \tag{A.25}$$

Applying Proposition 1 together with (A.25), we then conclude that $\|\hat{\theta}_n-\theta^*\|_2=\mathcal{O}(\delta_n)$ as $n\to\infty$.

□

A.6 Proof of Proposition 3 (Informal)

Proof.

Part i. We simply verify the conditions of Theorem 1, where f_n comes from a neural network estimate with the structure of Equation (9). Note that $\phi(\cdot)$ in (9) is a continuous function, and so is f_n. Moreover, due to the result from Cybenko (1989), when the target function f is a continuous function defined over a compact set, the output of an appropriately trained single-layer neural network uniformly converges to the target function. Therefore, the conditions of Theorem 1 are satisfied, and thus $\hat{\theta}_n \to \theta^*$.

Part ii. Knowing that the assumptions in Proposition 1 hold, we use its corresponding notation. Based on the assumptions of Proposition 1, Step 2 in the proof of Proposition 2 shows that f is Lipschitz continuous on an open ball $\mathbf{B}(\theta^*, \beta_0) \subseteq \mathbf{P}_\theta$, with Lipschitz factor L_1. Moreover, consider $\hat{\mathbf{S}} = \bar{\mathbf{B}}(\theta^*, \beta_0/2) \subset \mathbf{B}(\theta^*, \beta_0)$ and note that for sufficiently large n, $\hat{\theta}_n$ lies in $\hat{\mathbf{S}}$. Now, by applying Proposition 1, we conclude that:

$$\|\hat{\theta}_n - \theta^*\|_2 \leq \left\|\Lambda_n \left(f(\hat{\theta}_n) - f_n(\hat{\theta}_n)\right)\right\|_2 + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right) \tag{A.26}$$

$$\leq L_0 \sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2 + \mathcal{O}\left(\|\hat{\theta}_n - \theta^*\|_2^2\right). \tag{A.27}$$

Next, we investigate the bound on $\sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2$. Let f^r and f_n^r denote the r-th component functions of f and f_n, respectively. According to Barron (1994), the mean integrated squared error of the neural network estimator f_n^r for f^r is $\mathcal{O}(\sqrt{(K/n)\log n})$, provided that the number of nodes satisfies $N = \sqrt{n/(K \log n)}$. More formally, we have:

$$\frac{1}{\text{vol}(\mathbf{P}_\theta)} \int_{\mathbf{P}_\theta} |f^r(\theta) - f_n^r(\theta)|^2 \, d\theta = \mathcal{O}\left(\left(\frac{K \log n}{n}\right)^{\frac{1}{2}}\right). \tag{A.28}$$

The goal is to translate the bound in (A.28) to a bound on $\sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2$. To this end, we proceed as follows: The uniform boundedness of $\|\nabla f_n(\theta)\|_2$ across n and θ implies that there exists $L_2 > 0$ such that every function f_n is Lipschitz continuous with Lipschitz constant L_2. Therefore, letting $h_n(\theta) = |f^r(\theta) - f_n^r(\theta)|$, we conclude that $h_n(\theta)$ is Lipschitz continuous on $\hat{\mathbf{S}}$ with Lipschitz constant $L_1 + L_2$. Furthermore, due to the continuity of h_n and the compactness of $\hat{\mathbf{S}}$, there exist $\theta_{\min} = \arg\min_{\theta \in \hat{\mathbf{S}}} h_n(\theta)$ and $\theta_{\max} = \arg\max_{\theta \in \hat{\mathbf{S}}} h_n(\theta)$. Consequently, by the Lipschitz continuity of h_n on $\hat{\mathbf{S}}$, we have:

$$h_n(\theta_{\max}) \leq h_n(\theta_{\min}) + (L_1 + L_2) \|\theta_{\max} - \theta_{\min}\|_2 \leq h_n(\theta_{\min}) + (L_1 + L_2)\, \beta_0. \tag{A.29}$$

Recall that β_0 is the diameter of $\hat{\mathbf{S}}$. Moreover, the following inequality holds:

$$h_n^2(\theta_{\min}) \leq \frac{1}{\text{vol}(\hat{\mathbf{S}})} \int_{\hat{\mathbf{S}}} |f^r(\theta) - f_n^r(\theta)|^2 \, d\theta \leq \frac{1}{\text{vol}(\hat{\mathbf{S}})} \int_{\mathbf{P}_\theta} |f^r(\theta) - f_n^r(\theta)|^2 \, d\theta. \tag{A.30}$$

Now, (A.28) and (A.30) together imply that $h_n(\theta_{\min}) = \mathcal{O}(((K/n)\log n)^{1/4})$. Applying this to (A.29) yields the same bound on $h_n(\theta_{\max})$. Hence, $\sup_{\theta \in \hat{\mathbf{S}}} |f^r(\theta) - f_n^r(\theta)| = \mathcal{O}(((K/n)\log n)^{1/4})$.

Therefore, if we use M neural networks with $N = \sqrt{n/(K \log n)}$ nodes for each to estimate $f^1, \ldots, f^M$, we obtain the following bound for the overall error:

$$\sup_{\theta \in \hat{\mathbf{S}}} \|f(\theta) - f_n(\theta)\|_2 \leq \sqrt{M} \max_{1 \leq r \leq M} \sup_{\theta \in \hat{\mathbf{S}}} |f^r(\theta) - f_n^r(\theta)| = \mathcal{O}\left(\sqrt{M}\left(\frac{K \log n}{n}\right)^{1/4}\right). \tag{A.31}$$

Applying (A.31) to (A.27) then completes the argument.

□

B Corporate finance model: Implementation details

B.1 Construction of moments

We start with a COMPUSTAT extract over 1970-2019. We only keep firms that appear at least twice in the sample. We drop firms in the financial (SIC code 6) or regulated (SIC code 49) sectors. We also drop observations with total assets that are less than 10 million real 1982 dollars, or sales or book assets that grow by more than 200%. This results in a sample of 117,976 firm-year observations and 11,198 unique firms.

Using this COMPUSTAT sample, we compute seven baseline data moments, targeted in the baseline estimation, as follows: m_1, the average investment to capital ratio, is $\frac{\texttt{capx}}{l.\texttt{at}}$. m_2, the average profit to asset ratio, is $\frac{\texttt{oibdp}}{l.\texttt{at}}$. m_3, the average equity issuance to asset ratio, is computed net of repurchases: $\frac{\texttt{sstk}-\texttt{prstkc}}{l.\texttt{at}}$. m_4, mean net leverage, is $\frac{\texttt{dlc}+\texttt{dltt}\ -\texttt{che}}{\texttt{at}}$. m_5, the autocorrelation of investment rates, is measured by regressing $\frac{\texttt{capx}}{l.\texttt{at}}$ on its lag, with year fixed-effects. Last, m_6 and m_7, are the sample standard deviations of 1-year and 5 year log sales growth: $\log \texttt{sale} - \log \texttt{l.sale}$ and $\log \texttt{sale} - \log \texttt{l5.sale}$. All ratios are winsorized at the median +/- five times the interquartile range. We also remove firm fixed-effects from all the variables used in the empirical analysis, as the model does not feature any source of fully persistent heterogeneity across firms: for each variable, we subtract the within-firm average and add back the overall sample average. Table B.1, column 1, provides the values of these moments (together with their s.e.) in bold font.

Table B.1: Data and simulated moments (Corporate Finance Model)

		(1) Data	(2) True SMM	(3) Approximate	(4) Simulation
m_1	**mean(investment/assets)**	.0760 (.0007)	.0760	.0760	.0759
m_2	**mean(profit/assets)**	.1343 (.0011)	.1343	.1343	.1343
m_3	**mean(equity issuance/assets)**	.0158 (.0006)	.0159	.0158	.0164
m_4	**mean(leverage)**	.1049 (.0029)	.1050	.1049	.1035
m_5	**autocorr(investment/assets)**	.3754 (.0066)	.3758	.3754	.3869
m_6	**std(log sales growth)**	.2270 (.0017)	.2270	.2270	.2259
m_7	**std(log sales growth 5yr)**	.5851 (.0050)	.5851	.5851	.5890
m_8	var(investment/assets)	.0033 (.0001)	.0167	.0176	.0167
m_9	var(equity issuance/assets)	.0071 (.0002)	.0024	.0020	.0025
m_{10}	frequency(equity issuance)	.1178 (.0015)	.1534	.1565	.1576
m_{11}	coeff. regr. investment ratio on market/book	.0122 (.0005)	.3188	.3051	.3074
m_{12}	coeff. regr. net leverage on market/book	-0.0348 (.0018)	-0.0313	-0.0278	-0.0285
m_{13}	coeff. AR(1) regr. of profit/assets	.5210 (.0055)	.5357	.5445	.5433
m_{14}	resid std AR(1) regr. of profit/assets	.0728 (.0006)	.0286	.0285	.0285
m_{15}	var(leverage)	.0266 (.0004)	.0002	.0002	.0002
m_{16}	mean(dividend/assets)	.0267 (.0004)	.0492	.0505	.0498
m_{17}	var(dividend/assets)	.0013 (.0000)	.0054	.0052	.0053

Notes. The "Data" column reports moments in the data with standard errors in parenthesis. Column "True SMM" reports simulated moments using the true economic model $f(\theta)$ and parameters estimated using the true SMM. Column "Approximate" reports moments calculated using the benchmark neural net approximation f_n and the parameters estimated using the approximate SMM. Column "Simulation" reports moments calculated from the true economic model $f(\theta)$ but using parameter estimates from the approximate SMM. Targeted moments in the baseline SMM estimations are shown in bold font.

In addition to the seven baseline moments, we also compute ten additional moments for conducting the robustness analysis, using the same COMPUSTAT sample. These moments are: $m_8 =$ var(investment/assets), $m_9 =$ var(equity issuance/assets), $m_{10} =$ frequency(equity issuance), $m_{11} =$ coefficient of the regression

of investment ratio on market to book ratio, m_{12} = coefficient of the regression of net leverage on market to book ratio, m_{13} = coefficient of the AR(1) regression of profit ratio with year fixed-effects, m_{14} = residual std of the AR(1) regression of profit ratio with year fixed-effects, m_{15} = var(net leverage), m_{16} = mean(dividend/assets), m_{17} = var(dividend/assets). Table B.1, column 1, provides the values of these moments (together with their s.e.) in regular font.

B.2 Kernel approximation: Implementation

We use a kernel function that maps parameters into moments using the formula (6):

$$f_n(\theta) = \frac{\sum_{i=1}^{n} k_n(\theta, \theta_i) f(\theta_i)}{\sum_{i=1}^{n} k_n(\theta, \theta_i)}, \tag{B.1}$$

where the kernel itself is a Gaussian kernel such that:

$$k_n(\theta, \theta_i) = \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\left(\frac{\theta_k - \theta_{i,k}}{b_k}\right)^2\right)$$

where θ_k and $\theta_{i,k}$ are individual parameter number k of the vector of parameters, θ and θ_i, and b_k is the bandwidth corresponding to parameter number k. Following proposition 2, we set the bandwitdh to be:

$$b_k = \gamma\left(\overline{\theta_k} - \underline{\theta_k}\right)\left(\frac{1}{n}\right)^{\frac{1}{K}}$$

where $\left(\overline{\theta_k} - \underline{\theta_k}\right)$ is the range of parameter number k. n is the number of points in the training sample and K the number of parameters. The factor γ, common to all parameters, is used to fit the kernel approximation. We use $\gamma = 0.5$ in our application to get the best in-sample fit.

B.3 Neural net approximation: Implementation

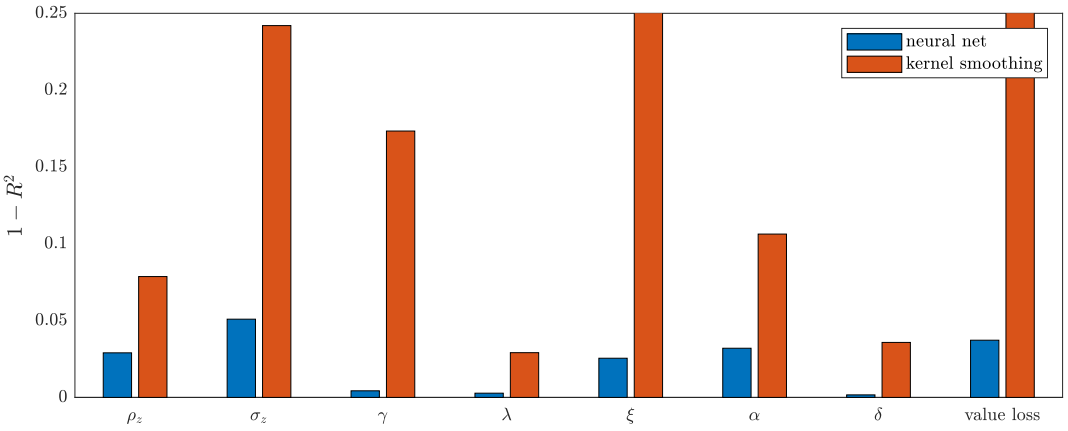
The neural net is a pyramidal MLP with 5 layers and 512, 256, 128, 64 and 32 nodes. The activation function is $ReLu$ at all layers. The loss function is the mean absolute error (which turns out to give a better in-sample fit than the MSE). The optimization algorithm is ADAM, a modern popular version of stochastic gradient descent with adaptive learning rate. The maximum number of epochs is 40, batch size is 256 points, initial learning rate is 0.005 (divided by 10 every 10 epochs). When training the neural net, we set aside 10% of the training sample as the validation and 10% as the test sets.

B.4 The superiority of the neural net fit

We compare the fit of neural net (NN) and Kernel smoothing. Figure B.1 reports the precision of our estimation method on the out-of-sample evaluation set, for each parameter separately, and for each parametrization of f_n. We report one minus the R^2 of true parameters regressed on estimated ones. A value of 0 means that estimated parameters are perfectly (and linearly) correlated with true ones.

First, Figure B.1 shows that overall our NN is much more accurate than the Kernel method. This is expected given that NNs converge faster than Kernel method, especially in high dimensions (see our propositions 2 and 3; see also Farrell et al. (2021) for deep NNs). Second, we see that the NN performs very well for all parameters, with an R^2 greater than 95% for all deep parameters. Therefore, we adopt the NN parameterization for f_n in conducting robustness checks.

Figure B.1: Out-of-sample performance, approximation method (Corporate Finance Model)

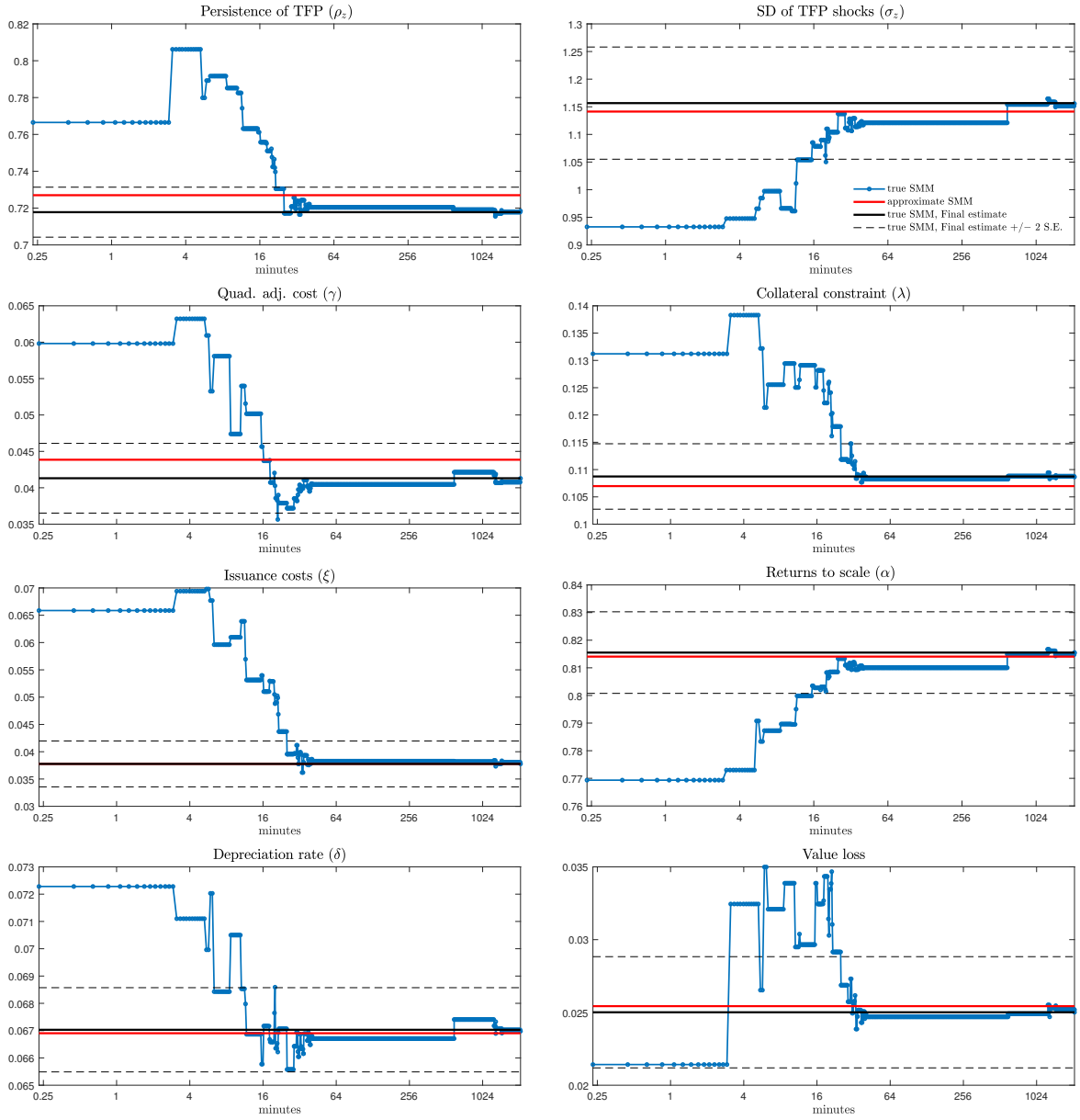


Notes. This figure reports a measure of the estimation error, on the out-of-sample evaluation set, from the approximate SMM for two approximations: Kernel smoothing and Neural net. For each one of the two approximations, and each one of the 7 parameters plus the value loss, we report $1 - R^2$, one minus the R^2 of a linear regression of the approximate moment estimate on the true parameter value. If the approximate moment estimate is exactly equal to the true value, this quantity is zero. When it is uncorrelated with the true value, it is 1.

B.5 Speed gains from the approximate SMM

We measure and compare the time for the true and approximate SMM. Figure B.2 shows that our approach is faster than the standard SMM by several orders of magnitude. The computing times we report exclude the simulation of the training sample, which is long but required for both estimations. For the true SMM, it takes at least an additional 20 minutes for the estimation to converge to its final value. In contrast, the approximation-based estimation converges in less than a second (provided the neural net has been estimated).

Figure B.2: SMM estimates, Convergence speed of the local optimization stage (Corporate Finance Model)



Notes. We report parameter values estimated as a function of time taken via two different SMM estimations. The blue line corresponds to the local optimization stage of the true SMM, i.e. the minimization of the distance of empirical moments to the true model $f(\theta)$. The optimization algorithm used in this case is Tiktak, using 50 starting points selected from a training set of 50,000 cases and Nelder-Mead algorithm for optimization per starting point with 200 max function iteration. The red line corresponds to the benchmark approximate SMM, which uses a neural net. The approximate estimation requires .9 seconds — hence the red line jumps to its final value at the origin. The black line corresponds to the true SMM estimate and the dashed lines represent the confidence interval (+/-2 standard errors) around it.

C Household finance model: Implementation details

C.1 Calibrated parameters

The parameters of income, stock market and social security are drawn from Catherine (2021) and summarized in Table C.1.

Table C.1: Pre-set Parameters of Life-cycle Model

p_z	.136	r	.02	p_s	.146
ρ_z	.967	φ_1	.1237	μ_s^-	-.245
$\overline{\mu_z^-}$	-.086	φ_2	-.0125	μ_s^+	.115
λ_{zl}	4.291	φ_0	-3.015	σ_{s1}	.077
σ_z^-	.562	t_0	23	σ_{s2}	.114
σ_z^+	.037	R	65	μ_l	.008
σ_η^-	.895	T	100	λ_{ls}	.161
σ_η^+	.089			σ_l	.017

Notes. This table shows the calibrated parameters used in the estimation of the life-cycle model, introduced in Section 5.1.

C.2 Construction of data moments

We compute the three baseline data moments, targeted in the baseline estimation, using the 1989–2016 waves of the triennial Survey of Consumer Finances (SCF). We restrict the sample to households whose head is between age 22 and 99 and have positive net worth. The three baseline moments are estimated as:

- m_1, the mean wealth, which is measured using the net worth variable (*networth*) from the SCF summary extract public data; note that to improve comparability across survey years, we scale wealth by the average wage income (*wageinc*) of each survey year
- m_2, the average participation rate, which is the share of households whose total holdings of stock (*equity*) is strictly positive
- m_3, the mean conditional equity share, which is the total holdings of stock (*equity*) divided by net worth, excluding vehicles (*vehic*), only computed across households with strictly positive holdings of stocks.

We also compute two additional moments using the same SCF sample, which we use in conducting the robustness analysis; m_4: the median wealth, and m_5: the median conditional equity share. Table C.2, column 1, reports the values of the constrcuted data moments (together with their s.e.).

Table C.2: Data and simulated moments (Household Finance Model)

		Data	True SMM	Approximate	Simulation
m_1	**mean(wealth)**	5.633 (0.028)	5.633	5.633	5.620
m_2	**participation rate**	0.557 (0.002)	0.557	0.557	0.528
m_3	**mean(cond. equity share)**	0.345 (0.003)	0.345	0.345	0.340
m_4	median(wealth)	1.996 (0.013)	2.919	2.908	2.936
m_5	median(cond. equity share)	0.224 (0.002)	0.269	0.273	0.259

Notes. This table reports the moments targeted in the baseline estimation, in bold fonts, and untargeted moments, representing median of statistics, in regular fonts. Column "Data" shows the empirical moments, with standard errors in parenthesis. Column "True SMM" corresponds to the simulated moments for the parameters using the true SMM estimation. Column "Approximate" show the approximate moments at the approximate parameter estimates. Column "Simulation" reports the true simulated moments at the approximate parameter estimates.

C.3 Building the training and evaluation samples

We restrict the set of parameters $\mathcal{P}$ for the household finance model to: $\gamma \in [1; 20]$, $\beta \in [.5; 1]$, $\Phi \in [0; .25]$.

To build the training sample, we first draw a Halton sequence of $n = 2,000$ parameters (we use a smaller training sample to explore the effect of sample size). For each draw θ_i, we simulate the model and compute the model-generated moments $f(\theta_i)$. We then remove 8 points which correspond to absurd observations (wealth higher than 1,000 times the national income). We end up with a sample of 1,992 observations for training the fitted moment functions. When training the neural net, we set aside 10% of this sample as the validation and 10% as the test sets.

To measure the performance of our approximate SMM, we also build a separate evaluation dataset starting with 200 additional uniform draws of parameters from $\mathcal{P}$ and their corresponding moments. We then drop 1 absurd draw. Also, like in the corporate finance model, we remove "badly identified" draws. We label a draw as "badly identified" when an implied s.e. from the classic formula $(J^\top W J)^{-1/2}$,[8] is more than 100 times larger than the s.e. of SMM estimate using the real-world data moments. We remove such draws and end up with 106 points in the final evaluation set.

[8] In this formula, $J = \nabla f(\theta_i)$ is the Jacobian matrix at the parameter draw θ_i and W is the inverse variance matrix, for the three baselie moments.

D Appendix Figures and Tables

Table D.1: Literature

Sub-field	Year	Reference	Comparative statics	Jacobian	AGS	N/A
Investment, capital structure, and financing	1992	Whited JF				✓
	2003	Love RFE				✓
	2005	Hennessy et al. JF				✓
	2007	Hennessy et al. JF		✓		
	2011	DeAngelo et al. JFE	✓			
		Lin et al. JFE				✓
	2013	Matvos RFS				✓
	2014	Nikolov et al. JF	✓			
	2016	Warusawitharana et al. RFS	✓			
		Li et al. RFS				✓
	2017	Bakke et al. JFE	✓			
		Gu JFE			✓	
	2018	Wu RFS		✓	✓	
	2019	Nikolov et al. JFE	✓			
	2021	Frank et al. RFS				✓
		Begenauet al. JFE				✓
	(Forthcoming)	Catherine et al. JF	✓	✓	✓	
Corporate governance	2009	Gayle et al. AER				✓
	2010	Taylor JF				✓
		Kang et al. JFE				✓
	2012	Coles et al. JFE		✓		
	2013	Taylor JFE				✓
	2017	Jung et al. JFE		✓		
	2018	Page JFE			✓	
	2022	Bertomeu JFE				✓
Bankruptcy	2016	Glover JFE				✓
Banking	2014	Schroth et al. JFE		✓		
Corporate control	2014	Dimopoulos et al. JFE	✓			
	2015	Albuquerque et al.JF	✓			
	2018	Li et al. JFE		✓		
		Wang JFE				✓
	2020	Wang JFE	✓			
Entrepreneurship	2020	Jones et al. AER		✓		
	2022	Ewens et al. JFE	✓			
		Catherine JFE	✓			
Household finance	2018	Pagel Econometrica				✓
	2019	Sun et al. JFE	✓			
	2020	Ameriks et al. JPE				✓
	2022	Catherine RFS				✓
Real estate	2015	Corbae et al. JPE				✓
	2017	Landvoigt RFS				✓
	2020	Oh et al. JFE				✓
	2021	Ghent JFE				✓